

---

# LEVERAGING ORTHOGRAPHIC INFORMATION TO IMPROVE MACHINE TRANSLATION OF UNDER-RESOURCED LANGUAGES

---

Insight SFI Research Centre for Data Analytics  
Data Science Institute  
College of Science and Engineering  
National University of Ireland Galway

A thesis submitted in partial fulfilment of the requirements for the  
degree of

Doctorate of Philosophy

for

Bharathi Raja Asoka Chakravarthi  
B.Tech., M.Sc.

Supervisors:

Dr. John Philip McCrae  
Dr. Mihael Arcan

Examination Committee:

Dr. Paul Buitelaar  
Prof. Kevin Scannell

Chair:

Prof. Mathieu d'Aquin

Galway, Ireland, 2020





# Declaration

I, Bharathi Raja Asoka Chakravarthi, do hereby declare that I carried out this doctoral thesis entitled “**LEVERAGING ORTHOGRAPHIC INFORMATION TO IMPROVE MACHINE TRANSLATION OF UNDER-RESOURCED LANGUAGES**” at the Insight SFI Research Centre for Data Analytics, Data Science Institute, College of Engineering and Informatics, National University of Ireland Galway. It has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Signature: \_\_\_\_\_

*July 24, 2020*

Bharathi Raja Asoka Chakravarthi



# Acknowledgements

The work described in this thesis has been supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight), 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 731015 (ELEXIS-European Lexical Infrastructure) and 825182 (Prêt-à-LLOD) and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages) and the Enterprise Ireland (EI) Innovation Partnership Programme under grant agreement No IP20180729, NURS - Neural Machine Translation for Under-Resourced Scenarios.

*July 24, 2020*

Bharathi Raja Asoka Chakravarthi



# Preface

This thesis presents Bharathi Raja Asoka Chakravarthi's work on machine translation for under-resourced languages. As the internet becomes increasingly available to the whole world's population, we are seeing an increase in the number and availability of languages as digital content. It is of course vital that communication is made between language communities and machine translation is an increasingly essential tool for achieving this. This thesis concentrates on the use of orthographic information, recognizing the research gap, that has been ignored by many researchers in Europe and North America, in that many languages are written in their own unique script. This is, of course, clearly the case with the Dravidian languages that are the main focus of study in this work. This thesis presents several key new breakthroughs: firstly, it was shown that recognising orthography can be a simple method for removing code-mixing steps and it was shown that the removal of code-mixing improves the quality of machine translation. Secondly, it was shown not only that converting different orthographies to single representation is of value for translation, but more interestingly that the granularity of the target representation matters significantly. Finally, the topic of alignment across orthographies was studied and it was shown that using linguistically-grounded methods combined with state-of-the-art deep learning methods can effectively align lexicons for even severely under-resourced languages. As such, this thesis provides compelling insights for researchers working on machine translation and under-resourced languages and will be of interest to many other researchers in these areas.

*May 1, 2020*

John P. McCrae



# Abstract

This thesis describes our improvement of word sense translation for under-resourced languages utilizing orthographic information with a particular focus on creating resources using machine translation. The first target of this thesis is cleaning the noisy corpus in the form of code-mixed content at word-level based on orthographic information to improve machine translation quality. Our results indicate that the proposed removing of code-mixed text based on orthography results in improvement for Dravidian languages. We then turn our interest to the usage of training data from closely-related languages. While languages within the same language family share many properties, many under-resourced languages are written in their own native script, which makes taking advantage of these language similarities difficult. We propose to alleviate the problem of different scripts by transcribing the native script into a common representation such as the Latin script or the International Phonetic Alphabet (IPA). We also show that our method could aid the creation or improvement of wordnets for under-resourced languages using machine translation. Further, we investigate bilingual lexicon induction using pre-trained monolingual word embeddings and orthographic information. We use existing resources such as IndoWordNet entries as a seed dictionary and test set for the under-resourced Dravidian languages. To take advantage of orthographic information, we propose to bring the related languages into a single script before creating word embeddings, and use the longest common subsequence to take advantage of cognate information. Our methods for under-resourced word sense translation of Dravidian languages outperformed state-of-the-art systems in terms of both automatic and manual evaluation.

Key words: under-resourced languages, machine translation, closely related languages, orthographic information, phonetic transcription



# Contents

<b>Declaration</b>	<b>i</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>Preface</b>	<b>v</b>
<b>Abstract</b>	<b>vii</b>
<b>List of figures</b>	<b>xi</b>
<b>List of tables</b>	<b>xv</b>
<b>I Introduction</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Motivations and Research Questions . . . . .	5
1.2 Outline . . . . .	8
1.3 Publications . . . . .	9
<b>II Background and Literature Review</b>	<b>11</b>
<b>2 Background</b>	<b>13</b>
2.1 Under-resourced Languages . . . . .	13
2.2 Dravidian Languages . . . . .	14
2.3 Orthographic Information . . . . .	16
2.4 Statistical Machine Translation (SMT) . . . . .	18
2.5 Neural Machine Translation (NMT) . . . . .	19
2.6 Bilingual Lexicon Induction (BLI) . . . . .	20
2.7 Automatic Evaluation . . . . .	22
2.8 Summary . . . . .	22
<b>3 Literature Review</b>	<b>23</b>
3.1 Orthographic Information in SMT . . . . .	23
3.2 Orthographic Information in NMT . . . . .	27

## Contents

---

3.3	Orthographic Information in Bilingual Lexicon Induction . . . . .	30
3.4	Related Work on Wordnet Creation . . . . .	31
3.5	Summary . . . . .	32
<b>III</b>	<b>Code Mixing</b>	<b>33</b>
<b>4</b>	<b>Improving Word Sense Translation by Removing Code-mixing</b>	<b>35</b>
4.1	Introduction . . . . .	35
4.2	Our Approach . . . . .	38
4.3	Results . . . . .	43
4.4	Discussion and Conclusion . . . . .	46
4.5	Summary . . . . .	46
<b>IV</b>	<b>Phonetic Transcription</b>	<b>47</b>
<b>5</b>	<b>Comparison of Different Orthography</b>	<b>49</b>
5.1	Introduction . . . . .	49
5.2	Our Approach . . . . .	51
5.3	Results . . . . .	54
5.4	Conclusion . . . . .	59
<b>6</b>	<b>Word Sense Translation using Multilingual NMT</b>	<b>61</b>
6.1	Introduction . . . . .	61
6.2	Our Approach . . . . .	62
6.3	Experimental Setting . . . . .	63
6.4	Results . . . . .	65
6.5	Conclusion . . . . .	67
<b>7</b>	<b>Multilingual Multimodal NMT utilising Phonetic Transcription</b>	<b>69</b>
7.1	Introduction . . . . .	69
7.2	Improving Multimodal NMT with Multilingual Approach . . . . .	72
7.3	Experimental Settings . . . . .	73
7.4	Results . . . . .	75
7.5	Conclusion . . . . .	77
7.6	Summary . . . . .	77
<b>V</b>	<b>Orthography in Bilingual Lexicon Induction</b>	<b>79</b>
<b>8</b>	<b>Bilingual Lexicon Induction across Orthographically-distinct Languages</b>	<b>81</b>
8.1	Introduction . . . . .	81
8.2	Our Approach . . . . .	82
8.3	Results and Discussion . . . . .	89

8.4 Conclusion . . . . .	91
<b>VI Conclusions and Future Work</b>	<b>93</b>
<b>9 Conclusions and Future work</b>	<b>95</b>
9.1 Conclusion and Research Questions Answered . . . . .	95
9.2 Contributions . . . . .	97
9.3 Future Work . . . . .	98
<b>Bibliography</b>	<b>119</b>



# List of Figures

2.1	Classification of world languages based on the availability of resources. . . . .	14
2.2	Examples of the multilingual NMT. . . . .	20
4.1	Examples of code-mixing in Tamil-English parallel corpus. In the first example the verb <i>loving</i> is code-mixed in Tamil. In second example the noun <i>GNOME</i> is code-mixed. . . . .	36
4.2	Work flow of creating a wordnet for Dravidian languages by cleaning up code-mixed corpora. . . . .	38
4.3	Examples of the before and after removal of code-mix from the corpora. . . . .	41
4.4	Examples of the manual evaluation of Tamil wordnet entries in comparison to the IndoWordNet (IWN). . . . .	45
5.1	Examples of phonetic transcription. . . . .	49
5.2	Example of closely related languages . . . . .	50
5.3	Orthographic representation of word <i>blue</i> in Tamil, Telugu and Kannada shown in native script, Latin script and IPA. . . . .	54
6.1	Our approach for wordnet gloss translation using multilingual NMT . . . . .	63
7.1	An example of multimodal MT with an image, a description in English sentence, and translation in Tamil. . . . .	69
7.2	Example of sentences with special tokens to indicate the source and target languages. . . . .	71
7.3	Example of sentence and image with candidate translation to choose. . . . .	75
7.4	Example showing improvement of translation quality and readability of the translation over baseline model. Errors are shown in red color. . . . .	76
7.5	Example showing translation with accurate transfer of important information. Errors are shown in red. . . . .	77
8.1	Example of cognate words in Dravidian language . . . . .	83



# List of Tables

4.1	Statistics of the parallel corpora used to train the translation systems. . . . .	39
4.2	Number of sentences (sent) and number of tokens (tok) removed from the original corpus. . . . .	41
4.3	Automatic translation evaluation of the of 1,000 randomly selected sentences in terms of the BLEU metric. . . . .	42
4.4	Results of Automatic evaluation of wordnet with IndoWordNet Precision at different level denoted by P@10 which means Precision at 10. . . . .	44
4.5	Manual Evaluation of wordnet creation for Tamil language compared with IndoWordNet (IWN) at precision at 10 presented in percentage. . . . .	45
5.1	Corpus statistics of the <b>complete corpus</b> (Collected from OPUS on August 2017) used for MT. (Tokens-En: Total number of tokens in the English side of parallel corpora. Tokens-Dr: Total number of tokens in the Dravidian language side of parallel corpora.) . . . . .	52
5.2	Corpus statistics of the <b>multi-parallel corpus</b> used for MT. (Tokens-En: Total number of tokens in the English side of parallel corpora. Tokens-Dr: Total number of tokens in the Dravidian language side of parallel corpora.) . . . . .	52
5.3	Cosine similarity of the transliteration of the languages under study at character level using the <b>complete corpus</b> . . . . .	54
5.4	Cosine similarity of the transliteration of the languages under study at character level using the <b>multi-parallel corpus</b> . . . . .	54
5.5	BLEU (B), METEOR (M) and chrF (C) scores are illustrated for systems trained with the native script, Latin script and IPA. The native script is different for Tamil, Telugu and Kannada. Latin script and IPA are common script representations. Best results for each system are shown in bold. . . . .	55
5.6	BLEU (B), METEOR (M) and chrF (C) scores are shown for systems trained with the native script, Latin script and IPA for <b>multi-parallel corpora with different evaluation set</b> . The native script is different for Tamil, Telugu and Kannada. Latin script and IPA are common script representations. Best results for each system are shown in bold. . . . .	56
5.7	Manual evaluation results of 50 sentences from testset generated by different systems for Tamil. . . . .	58

## List of Tables

---

6.1	Statistics of the parallel corpora used to train the multilingual translation systems.	64
6.2	Number of sentences (sent) and number of tokens (tok) removed from the original corpus. . . . .	65
6.3	Results of Automatic evaluation of translated wordnet with IndoWordNet Precision at different level denoted by P@10 which means Precision at 10. B-Baseline original corpus, NC- Non-code mixed, MNMT-Multilingual Neural Machine Translation, NCT-MNMT Non Code-mixed Transliteration Multilingual Neural Machine Translation . . . . .	66
6.4	Manual evaluation of wordnet creation for Tamil language compared with IndoWordNet (IWN) at precision at 10 presented in percentage. B-Baseline original corpus, NC- Non-code mixed, MNMT-Multilingual Neural Machine Translation, NCT-MNMT Multilingual Neural Machine Translation . . . . .	67
7.1	Statistics of the parallel corpora used to train the general domain translation systems. sent: Number of sentences, s-tokens: Number of source tokens, and t-tokens: Number of target tokens. . . . .	74
7.2	Results of general domain SMT and NMT translation systems on general domain evaluation set . . . . .	74
7.3	Results are expressed in BLEU score: Baseline is Multimodal NMT, MMNMT is trained on native script, and MMNMT-T is trained utilizing phonetic transcription.	76
8.1	Examples from training set for comparison of NLCS and NL . . . . .	86
8.2	Comparison of NLCS and NL between languages for example of cognate words	87
8.3	Number of entries in the initial bilingual lexicons used as a seed dictionary for the experiment . . . . .	88
8.4	Number of entries in the test set for the experiment . . . . .	88
8.5	Number of sentences and number of tokens extracted from wikimedia.org for Dravidian languages . . . . .	89
8.6	Performance comparison of bilingual lexicon induction on test data for Dravidian languages. Translation accuracy is represented in percentage. N: native script, T: transliteration, seed-dic: seed dictionary. . . . .	90

# Introduction **Part I**



# 1 Introduction

Natural Language Processing (NLP) plays a significant role in keeping languages alive and the development of languages in the era of digital devices (Karakanta et al. 2018). One of the sub-parts of NLP is Machine Translation (MT) which has long been the most promising application of Artificial Intelligence (AI). MT has been shown to increase access to information through the native language of the speakers in many cases; one such important case is the spread of vital information during crisis or emergency in less common languages (Lewis et al.; Neubig and Hu 2011; 2018). Recently translation accuracy has increased, and commercial systems for MT have gained popularity. However, some of the less common languages of the world do not enjoy this availability of resources. These under-resourced languages lack essential linguistics resources, e.g. corpora, Part-Of-Speech (POS) taggers, or wordnets. This is more critical for MT since most common MT systems require large amounts of high-quality sentence aligned parallel corpora and other resources. Thus there is a need to develop a methodology to create MT systems for under-resourced languages. This research focuses on utilising orthographic information to improve word sense translation for Dravidian languages into or out of English. We apply our methods to improve and extend wordnets for the Dravidian languages, which are severely under-resourced. We report the evaluation results of the generated translations in terms of BLEU, METEOR, and chrF, and wordnet senses in terms of precision at different levels. In addition to that, we carried out a manual evaluation of the translations for the Tamil language, where we demonstrate that our approach can aid in improving wordnet resources for under-resourced languages.

The first target of this research is removing code-mixed content at word level. Word-level language identification of this corpus was made based on orthographic information. A clean data set was created from the available corpora for English-Tamil, English-Telugu, and English-Kannada. In this part of the research, we showed the challenges in building a wordnet for under-resourced languages and showed that our method could aid the creation or improvement of wordnets for under-resourced languages using MT. We experimented with cleaned data to create statistical machine translation (SMT) systems for three Dravidian languages into or out of English and used uncleaned data as a baseline. Our results indicated that the

proposed removing of code-mixed text based on orthography from the corpus results in gains for the wordnet entries with limited data.

Under-resourced languages are a significant challenge for statistical approaches to MT, but recently it has been shown that the usage of training data from closely-related languages can improve MT quality of these languages (Popović and Ljubešić; Abe et al. 2014; 2018). While languages within the same language family share many properties, many under-resourced languages are written in their own native script, which makes taking advantage of these language similarities difficult. In the second part of this research, we propose to alleviate the problem of different scripts by transcribing the native script into a common representation, i.e. the Latin script or International Phonetic Alphabet (IPA). We compare the difference between coarse-grained transliteration of the Latin script and fine-grained IPA transliteration. We performed experiments on the language pairs English-Tamil, English-Telugu, and English-Kannada translation task. Our results show improvements in terms of the BLEU, METEOR and chrF scores from transliteration and we find that the transliteration into the Latin script outperforms the fine-grained IPA transcription. We then trained a multilingual NMT systems on closely related languages utilising transliteration for Dravidian language corpora without code-mixing. We used this system to translate the senses in the Princeton WordNet which shows improvement in terms of precision. We also trained multilingual multimodal NMT utilising transliteration of this corpus and we have shown better results compared to without transliteration. We evaluate the generated wordnet sense in terms of precision. In comparison to the first part of the research, we show improvement in terms of precision.

Further, we investigate Bilingual Lexicon Induction (BLI) using pre-trained monolingual word embeddings and orthographic information. This work investigates the use of orthographic information for BLI between Dravidian languages, namely Tamil, Telugu, Kannada, and Malayalam, which use different scripts. Recent state-of-the-art approaches (Irvine and Callison-Burch 2017) use pre-trained monolingual word embeddings to create a bilingual lexicon using a supervised or semi-supervised approach. However, they require cross-lingual information in terms of seed dictionaries to train the model to find the linear transformation in case of supervised and semi-supervised approaches. Especially in the case of low-resourced languages, the seed dictionaries are not readily available. To take advantage of orthographic information and cognates in Dravidian languages, we bring the related languages into a single script. Previous approaches have used linguistically unsound 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.

### 1.1 Motivations and Research Questions

In this work, we investigate SMT, NMT, and BLL, which can exploit orthographic information from available resources to create MT of under-resourced languages. We study the effect of removing foreign words (code-mixing) based on orthography in Research Question 1. In Research Question 2, we convert the languages into a single script for the languages which use different scripts to take advantage of the phonetic similarities between closely-related languages. Specifically, we group languages based on language families. We utilise the Multilingual NMT (Firat et al. 2017), where multiple sources and target languages are trained simultaneously without changing the architecture. Research Question 3 studies how orthographic information can be utilised in BLL. We outline the following research questions (RQ), which are addressed in thesis.

- RQ1:** Does removing foreign words (code-mix) based on orthography improve word sense translation quality ?
- RQ2:** Does phonetic transcription of parallel corpora help to take advantage of resources from closely related language to improve translation quality?
- RQ3:** Does orthographic information help to improve embedding-based bilingual lexicon induction for closely-related languages?

#### 1.1.1 Code-Mixing

**RQ1:** Does removing foreign words (code-mix) based on orthography improve word sense translation quality?

Code-switching or code-mixing is mixing of words, phrases, and sentences from two or more languages within the same sentence or between sentences (Ayeomoni; Solorio and Liu; Diab et al.; Yoder et al. 2006; 2008; 2014; 2017). In this section, we show the experiments done to explore RQ1. In our publication Chakravarthi et al. (2018), we describe the effort towards generating and improving wordnets (Miller; Vossen; Bhattacharyya 1995; 1997; 2010) for Dravidian languages such as Tamil (ISO 639-3: tam), Telugu (ISO 639-3: tel) and Kannada (ISO 639-3: kan). We used available parallel corpora from multiple sources, such as OPUS (Tiedemann and Nygaard; Tiedemann 2004; 2012) to create an MT system to translate the wordnet senses in the Princeton WordNet into the above mentioned under-resourced languages. The parallel corpora from OPUS had many code-switching points in the corpora. The handling of code-mixing by removing them based on the orthography shown to improve the quality of the results outperforming the baseline results of wordnet entries. Thus, we believe that the method presented in this work is also applicable to resource creation of under-resourced languages with minimum effort.

Publications: Research Question 1

- Bharathi Raja Chakravarthi, Mihael Arcan and John P. McCrae “[Improving Wordnets for Under-Resourced Languages Using Machine Translation.](#)” In the proceedings of the 9th Global WordNet Conference, (2018).

### 1.1.2 Phonetic Transcription

**RQ2:** Does phonetic transcription of parallel corpora help to take advantage of resources from closely related language to improve translation quality?

In **RQ1**, we have shown that orthographic information can be used to improve the corpus quality consequently increasing the quality of word sense translation for under-resourced languages. However, under-resourced languages are a significant challenge for statistical approaches to MT, and recently it has been shown that the usage of training data from closely-related languages can improve MT quality of these languages (Popović and Ljubešić; Abe et al. 2014; 2018). Closely-related languages refer to languages that share similar lexical and structural properties due to sharing a common ancestor (Popović et al. 2016). Frequently, languages in contact with other language or closely-related languages like the Dravidian, Indo-Aryan, and Slavic family share words from a common root (*cognates*), which are highly semantically and phonologically similar. While languages within the same language family share many properties, many under-resourced languages are written in their own native script, which makes taking advantage of these language similarities difficult. We use phonetic transcription of corpus to overcome these restrictions.

Phonetic transcription is a method for writing the language in another script keeping the phonemic units intact. It is extensively used in speech processing research, text-to-speech, and speech database construction. Phonetic transcription into a single script has the advantage of collecting similar words at the phoneme level. In this research question, we propose to alleviate the problem of different scripts by transcribing the native script into a common representation, i.e. the Latin script or the IPA. We compare the difference between coarse-grained transliteration of the Latin script and fine-grained IPA transliteration. We performed experiments on the language pairs English-Tamil, English-Telugu, and English-Kannada translation task. Our results show improvements in terms of the BLEU, METEOR and chrF scores from transliteration and we find that the transliteration into the Latin script outperforms the fine-grained IPA transcription (Chakravarthi et al. 2019a). We trained a multilingual NMT systems on closely related languages utilising phonetic transcript (**RQ2**) for Dravidian language corpora without code-mixing (**RQ1**). We used this system to translate the senses in the Princeton WordNet which shows improvement in terms of precision (Chakravarthi et al. 2019b). We also trained multilingual multimodal NMT utilising transliteration of a corpus and shown better results compared to without transliteration (Chakravarthi et al. 2019c).

Publications: Research Question 2

- Bharathi Raja Chakravarthi, Mihael Arcan and John P. McCrae “[Comparison of Different](#)

[Orthographies for Machine Translation of Under-resourced Dravidian Languages.](#)” In proceedings of the 2nd Conference on Language, Data and Knowledge, Leipzig, Germany (2019).

- Bharathi Raja Chakravarthi, Mihael Arcan and John P. McCrae [“WordNet Gloss Translation for Under-resourced Languages using Multilingual Neural Machine Translation.”](#) In proceedings of Second Workshop on Multilingualism at the intersection of Knowledge Bases and Machine Translation – co-located with MT-Summit, Dublin, Ireland (2019).
- Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, S Sridevy, Mihael Arcan, Manel Zarrouk, and John P. McCrae [“Multilingual Multimodal Machine Translation for Dravidian Languages utilising Phonetic Transcription.”](#) In proceedings of Second Workshop on Technologies for MT of Low Resource Languages – co-located with MT-Summit, Dublin, Ireland (2019).

### 1.1.3 Embedding Based Bilingual Lexicon Induction utilising orthography

While **RQ1** and **RQ2** are based on orthographic information in MT scenario, we then switch our attention to orthographic information in the bilingual lexicon induction. As a matter of fact, the bilingual lexicon can be used to improve MT, when training data is insufficient ([Chu et al. 2014](#)).

**RQ3:** Does orthographic information help to improve embedding-based bilingual lexicon induction for closely-related languages?

A bilingual lexicon provides valuable information for semantic equivalence of words across languages. Building bilingual lexicons from monolingual corpora could benefit under-resourced languages which lack parallel corpora. These monolingual corpora can range from comparable corpora to completely unrelated documents. In this research question, we present an approach to building an embedding based multilingual lexicon leveraging orthographic similarity. This research question investigates the utilisation of orthographic information for bilingual lexicon induction between closely related languages in Dravidian languages, namely Tamil, Telugu, Kannada, and Malayalam, which use different scripts. Recent state-of-the-art approaches ([Irvine and Callison-Burch 2017](#)) use pre-trained monolingual word embeddings to create a bilingual lexicon using a supervised or semi-supervised approach. However, they require cross-lingual information in terms of seed dictionaries to train the model to find the linear transformation in case of supervised and semi-supervised approaches. Especially in the case of low-resourced languages, the seed dictionaries are not readily available.

Cognate have a similar meaning and similarity in orthography based on etymological relationship ([Kondrak et al. 2003](#)). 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. As it has been shown that even a little seed information can improve the performance over fully unsupervised methods (Patra et al. 2019), the usage of such alignments is likely to improve performance. To take advantage of orthographic information and cognates, we bring the related languages into a single script by phonetic transcription. It was shown in **RQ2** that phonetically transcribing the corpus into single script improves MT results. We also show that contrary to state-of-the-art approaches which have used linguistically unsound 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.

### 1.2 Outline

In this thesis, we address word sense translation for under-resourced languages in three scenarios: removing code-mixing, phonetic transcription in multilingual NMT, and BLI. Overall, the goal is to improve word sense translation quality by using orthographic information. This comprises of six parts including the current introductory **Part I** which includes Chapter 1, explains the structure of the thesis.

The **Part II** consists of Chapter 2 and Chapter 3. In Chapter 2, we provide background information about under-resourced languages and orthographic information in MT. Related work is presented in the Chapter 3. At the time of completing the thesis, **Part II** has been submitted to the Machine Translation Journal and is under-review.

In **Part III**, we study the code-mixing in available corpora for MT. In our work, we remove the code-mixing text based on the orthography. We show that our approach can improve word sense translation quality, and it can help to create wordnet in Chapter 4.

In **Part IV**, we move our attention to phonetic transcription in multilingual NMT. In Chapter 5, we present our comparative analysis with and without phonetic transcription in multilingual NMT. Then, we improve the wordnet creation with the phonetic transcription, closely related languages and multilingual NMT in Chapter 6. We also show that phonetic transcription can also help in multimodal multilingual NMT in Chapter 7.

In **Part V**, we focused on bilingual lexicon induction utilising orthographic information explained in Chapter 8. At the time of completing the thesis, **Part V** has been submitted to the 28th International Conference on Computational Linguistics (COLING 2020) and is under-review.

The last **Part VI** explains how this research contributes to our field and concludes the thesis with some avenues for future work. We conclude in Chapter 9 with a summary of our work and contributions of the thesis. Finally, we present avenues for future work .

In summary, all of our proposed approaches and experiments are presented in Part III, Part IV, and Part V, which are related to **RQ1**, **RQ2** and **RQ3**, respectively.

## 1.3 Publications

Other publications during this thesis period are listed below:

- Prakash Ranjan, Bharathi Raja, Ruba Priyadharshini, and Rakesh Chandra Balabantaray. “A Comparative Study on Code-Mixed Data of Indian Social Media vs Formal Text.” In Proceedings of the 2nd International Conference on Contemporary Computing and Informatics (IC3I 2016), pages 608–611, Noida, India, 2016. IEEE
- Daniel Torregrosa, Nivranshu Pasricha, Maraim Masoud, Bharathi Raja Chakravarthi, Juan Alonso, Noe Casas, and Mihael Arcan, “Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models.” In Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, 2019
- Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti and John P. McCrae, “Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 68-72.
- Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly and John P. McCrae, “A Survey of Current Datasets for Code-Switching Research,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 136-141.
- Pranav Verma, Bharath Sudharsan, Bharathi Raja Chakravarthi, Colm O’Riordan and Seamus Hill, “Unsupervised Method to Analyse Playing Styles of EPL Teams using Ball Possession-position Data,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 58-64.
- Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly and John P. McCrae “A Sentiment Analysis Dataset for Code-Mixed Malayalam-English,” In Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020), co-located with 12th Edition of Language Resources and Evaluation Conference, Marseille, France, May, 2020.
- Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini and John P. McCrae “Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text,” In Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020), co-located with 12th Edition of Language Resources and Evaluation Conference, Marseille, France, May, 2020.
- Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae “A Comparative Study of Different State-of-

the-Art Hate Speech Detection Methods for Hindi-English Code-Mixed Data,” In Proceedings of the TRAC-2: The Second Workshop on Trolling, Aggression & Cyberbullying, co-located with 12th Edition of Language Resources and Evaluation Conference, Marseille, France, May, 2020.

- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Paul Buitelaar “**Multi-modal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text,**” In Proceedings of the TRAC-2: The Second Workshop on Trolling, Aggression & Cyberbullying, co-located with 12th Edition of Language Resources and Evaluation Conference, Marseille, France, May, 2020.
- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Paul Buitelaar “**A Dataset for Classification of Tamil Memes,**” In Proceedings of the 5th Workshop on Indian Language Data Resource and Evaluation (WILDRE-5), co-located with 12th Edition of Language Resources and Evaluation Conference, Marseille, France, May, 2020.
- Md. Rezaul Karim, Bharathi Raja Chakravarthi, John P. McCrae, and Michael Cochez “**Classification Benchmarks for Under-resourced Bengali Language based on Multi-channel Convolutional-LSTM Network,**” submitted to Journal of Natural Language Engineering.
- Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan and John P. McCrae “A Survey of Orthographic Information in Machine Translation,” submitted to Journal of Machine Translation.
- Koustava Goswami, Priya Rani, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae “GenMA at SemEval-2020 Task 9: Generative Morphemes with an Attention Model for Sentiment Analysis in Code-Mixed Text,” In Proceedings of the 14th International Workshop on Semantic Evaluation, co-located with 28th International Conference on Computational Linguistics (COLING’2020), Barcelona, Spain, Dec, 2020.

# **Background and Literature Review Part II**



## 2 Background

In this section, we explain the necessary background information to follow the thesis, why we consider a language as under-resourced, orthographic information for MT, Dravidian languages, and different types of machine translation.

### 2.1 Under-resourced Languages

Worldwide, there are around 7,000 languages ([Abney and Bird; Hauksdóttir 2010; 2014](#)). However, most of the machine-readable data and natural language applications are available for very few popular languages, such as Chinese, English, French, or German. For other languages, resources are scarcely available and, for some languages, not at all. Setting aside Bible translations and linguistic research, probably more than half of these languages do not even have a writing system ([Maxwell and Hughes; Krauwer 2006; 2003](#)), or are not encoded in major schemes such as Unicode. Due to the unavailability of digital resources, many of these languages may go extinct. For any language that gets lost, we lose connection with the culture of the people and language characteristics.

[Alegria et al. \(2011\)](#) proposed six levels of language typology to develop language technologies that could be useful for several hundred languages. This classifies the world's languages based on the availability of Internet resources for each language. According to the study, the term resource-poor or under-resourced is relative and also depends on the year. The first level is the most resourced languages; the second level is languages amongst the top 10 languages most frequently used on the web. The third level is languages which have some form of resources in human language resources (HLT). The fourth level considers languages which have any lexical resources. Languages that have a writing system but are not in digital form are in the fifth level. The last level is significantly large, including oral languages which do not have a writing system of its own. We follow this approach to define the term under-resourced languages in terms of machine translation by taking the languages in the third and fourth level. Languages that lack extensive parallel corpora are considered as under-resourced or low-resourced languages ([Jimerson and Prud'hommeaux 2018](#)) for this thesis.

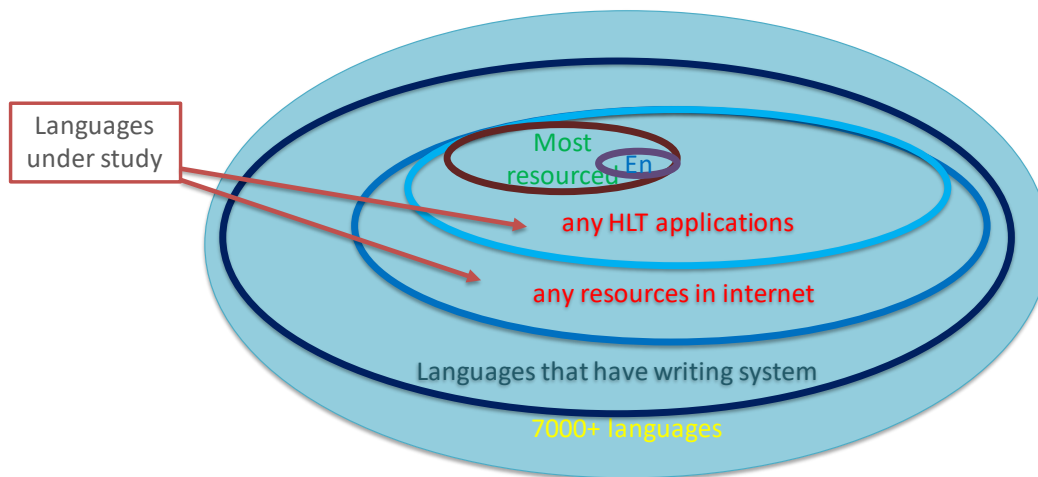


Figure 2.1 – Classification of world languages based on the availability of resources.

Languages that seek to survive in modern society need NLP, which requires a vast amount of data and linguistic knowledge to create new language technology tools for these languages. Mainly, it is a big challenge to develop MT systems for these languages due to the scarcity of data, specifically, sentence aligned data (parallel corpora) in large amounts to train MT systems. For example, Irish, Scottish Gaelic, Manx or Tamil, Telugu, and Kannada, belonging to the Goidelic and the Dravidian languages respectively, are considered as under-resourced languages due to scarcely available machine-readable resources (Alegria et al. 2011).

## 2.2 Dravidian Languages

Dravidian languages (Kumar et al. 2017) are spoken in the south of India by 215 million people. The Dravidian languages are divided into four groups: South, South-Central, Central, and North groups. Dravidian morphology is agglutinating and exclusively suffixal. Words are built from small elements called morphemes. Two broad classes of morphemes are stems and affixes. Words are made up of morphemes concatenated based on the grammar of the language.

The study of Dravidian languages suggests that these languages formed a single language dur-

ing 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. 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 the other language without being translated (Burrow and Emeneau 1961). The other languages in the Dravidian family still share many root words.

The four major literary Dravidian languages are Tamil, Telugu, Malayalam, and Kannada. Tamil, Malayalam, and Kannada fall under the South Dravidian subgroup, whereas Telugu belongs to the South Central Dravidian subgroup (Vikram and Urs 2007). All four languages have official status according to the Government of India and use their unique script. Outside India, Tamil also has official status in Sri Lanka and Singapore. However, the language resources for these languages are under-developed. These four Dravidian languages fall under the third and fourth category of Alegria et al. (2011) classification since these languages have some form of NLP tools and corpora on the Internet.

The Tamil script evolved from the Brahmi script, Vatteluttu alphabet, and Chola-Pallava script. It has 12 vowels, 18 consonants, and 1 *aytam* (voiceless velar fricative). The Telugu script is also a descendant of the Southern Brahmi script and has 16 vowels, 3 vowel modifiers, and 41 consonants. The Kannada script has 14 vowels, 34 consonants, and 2 *yogavahakas* (part-vowel, part-consonant). The Kannada and Telugu scripts are most similar, and are often considered as regional variants. The Kannada script is used to write other under-resourced languages like Tulu, Konkani, and Sankethi. Since Telugu and Kannada are influenced by Sanskrit grammar, the number of characters is higher than in the Tamil language. The Malayalam languages alphabet set contains 15 vowel letters, 42 consonant letters, and some other symbols. It is based on the Vatteluttu script that evolved from old Vatteluttu and Grantha script to write loan words. In contrast to Tamil, Malayalam, Kannada, and Telugu inherit some of the affixes from Sanskrit (Prakash and Joshi; Kumar et al. 1995; 2017).

Although they have nearly the same number of consonants and vowels, their orthographies differ due to historical reasons and whether they adopted the Sanskrit tradition or not (Bhanuprasad and Svenson 2008). Each of these has been assigned a unique block in Unicode, and thus from an MT perspective are completely distinct. To improve access to and production of information for monolingual speakers of Dravidian languages, it is necessary to have an MT system from and to English. However, Dravidian languages are under-resourced languages and thus lack the parallel corpus needed to train an MT system. In this thesis, we perform experiments on Tamil (ISO 639-3: tam), Telugu (ISO 639-3: tel), Malayalam (ISO 639-3: mal) and Kannada (ISO 639-3: kan).

### 2.3 Orthographic Information

Humans are endowed with a language faculty that is determined by biological and genetical development. However, it is not true of the written form of the language, which is the visual representation of the natural and genetically determined spoken form. With the development of orthography, humans have not only overcome limitations with human short term memory, and brain storage capacity, but also this development allows communication through space and time (Fromkin et al. 2018). Orthography is a linguistic factor of mutual intelligibility which may facilitate or impede inter-comprehension (Fischer et al. 2016).

The orthographic information of languages not only represents the information of the language but also the psychological representation of the world of the users. Chinese orthography is unique in the sense that it uses a logographic writing system. In such a system, each Chinese character carries visual patterns along with rich linguistic information. These characters are visualised in a square space, which depends on the number of strokes a character has. Each character can be decomposed in two parts. *Radicals*, which carry those semantic meaning, whereby the other part tell as about the pronunciation. According to Shuo WenJie Zi<sup>1</sup>, the new Chinese characters consist of 540 radicals but only 214 are still used in modern Chinese (Min et al. 2004). Problem arises when the decomposition strategy does not comply with some of the characters. On the other hand, other Asian languages such as Korean and Japanese, have two different writing systems. Modern-day Korean uses the Hangul orthography, which is part of the syllabic writing system, and the other is known as Hanja, which uses classical Chinese characters.

Like the history of writing in Korea, Japan has two writing systems, Kana and Kanji, where Kanji is imported from Classical Chinese characters, and Kana represents sounds where each kana character is a syllable. As both Korean and Japanese are very different from Chinese, these problems also posed great difficulty in the field of translation and transliteration. Irrespective of all the differences and challenges, these three Asian languages share common properties which could be significant advantages in MT.

Closely related languages share similar morphological, syntactical, orthographic properties and sometimes orthographic similarity can be seen from two primary sources. The first one is based on the genetic relationship between languages such as based on language families, Germanic, Slavic, Gaelic and Indo-Aryan languages. The second one is based on the contact though geographical area such as Indo-Aryan and Dravidian languages in the Indian subcontinent (Kunchukuttan et al. 2018). Two languages maintain orthographic similarity only when these languages have the following properties: overlapping phonemes, mutually compatible orthographic systems and similar grapheme to phoneme mapping.

The widespread and underlying problem for MT systems is variations in orthographic conventions. Two languages written in two different orthographies lead to errors in MT outputs

---

<sup>1</sup>[https://en.wikipedia.org/wiki/Shuowen\\_Jiezi](https://en.wikipedia.org/wiki/Shuowen_Jiezi)

since these systems cannot capture the similarity between these languages. Orthographic information can also be used to improve machine translation.

### **Spelling and Typographical Errors**

Spelling or typographical errors are to be handled very carefully in MT tasks as even minor spelling errors could generate out of vocabulary errors in the training corpus. The source and the target languages highly influenced the methodology used to correct orthographic errors.

### **True-casing and Capitalization**

The process of restoring case information to badly cased or not cased text is true-casing ([Lita et al. 2003](#)). To avoid orthographical errors, a popular method is to lower-case all words, especially in statistical machine translation. This method allows the system to avoid the mismatching of the same words, which seems different due to differences in casing. Thus converting all the text into lower-case is one of the methods to avoid these errors. In most MT systems, both pre-processing and post-processing is carried out. Post-processing of the text involves converting all the lower case to its original case form and generating the proper surface forms.

### **Tokenization and Detokenization**

The process of splitting text into smaller elements is known as tokenization. Tokenization can be done at different levels depending on the source and the target language as well as the goal that we want to achieve. It also includes the processing of the signs and symbols used in the text such as hyphens, apostrophes, punctuation marks, and numbers to make the text more accessible for further steps in MT. Like normalisation, tokenization also helps in reducing language data sparsity.

Detokenization is the process of combining all the tokens to the correct form before releasing the main output. Tokenization and detokenization are not linked directly to orthographic correction; rather, they are more about morphological linking and correction, especially towards morphologically abundant languages like Irish and Arabic ([Guzman et al. 2016](#)). Orthography plays a significant role in tokenization and detokenizations as each orthography has different rules on how to tokenise and detokenize.

### **Transliteration**

Transliteration is the conversion of the text from one orthography to another without any phonological changes. The best example of transliteration is named entities and generic words ([Kumaran and Kellner 2007](#)). Data collected from social media are frequently transliterated and contain errors; thus, using these data for building an MT system for resource-poor lan-

languages causes errors. One of the primary forms that have a high chance of transliteration is cognates. Cognates are words from different languages derived from the same root. Though MT has progressed a lot recently, however the method to the transliteration problem has changed from a language-independent manner to cognate prediction when translating between closely related languages. Transliteration of cognates has shown to improve results for under-resourced languages.

### Code-Mixing

Code-mixing is a phenomenon which occurs commonly in most multilingual societies where the speaker or writer alternates between more than one language in a sentence (Ayeomoni; Ranjan et al. 2006; 2016). Speakers whose first language uses a non-Latin script are writing using the Latin script for accessibility of modern digital devices keyboard (phonetic typing) which also increases the probability of code-mixing (Das and Gambäck 2013). This is similar for other languages where the script of official languages is used to write the local language.

Most of the corpora for under-resourced languages came from the publicly available parallel corpora which were created by voluntary annotators or aligned automatically. The translation of technical documents such as KDE, GNOME, and Ubuntu translations have code-mixed data since some of the technical terms may not be known to voluntary annotators for translation. Code-mixing in the OpenSubtitles corpus is due to bilingual and historical reasons of native speakers (Chanda et al.; Parshad et al. 2016; 2016). Different combinations of languages may occur while code-mixing, for example, German-Italian and French-Italian in Switzerland, Hindi-Telugu in state of Telangana, India, Taiwanese Hokkien-Mandarin Chinese in Taiwan (Chan et al. 2009) as a result, code-mixing of the script are also possible from a voluntary annotated corpus. This poses another challenge for MT.

## 2.4 Statistical Machine Translation (SMT)

Statistical Machine Translation (SMT) (Brown et al.; Koehn 1993; 2005) is one of the Corpus Based Machine Translation (CBMT) based approaches. SMT systems assume that we have a set of example translations  $(S^{(k)}, T^{(k)})$  for  $k = 1 \dots n$ , where  $S^{(k)}$  is the  $k^{th}$  source sentence,  $T^{(k)}$  is the  $k^{th}$  target sentence which is the translation of  $S^{(k)}$  in the corpus. SMT systems try to maximize the conditional probability  $p(t|s)$  of target sentence  $t$  given a source sentence  $s$  by maximizing separately a language model  $p(t)$  and the inverse translation model  $p(s|t)$ . A language model assigns a probability  $p(t)$  for any sentence  $t$  and translation model assigns a conditional probability  $p(s|t)$  to source / target pair of sentence (Wang and Waibel 1997). By Bayes' rule

$$p(t|s) \propto p(t)p(s|t) \tag{2.1}$$

This decomposition into a translation and a language model improves the fluency of generated texts by making full use of available corpora. The language model is not only meant to ensure a fluent output, but also supports difficult decisions about word order and word translation (Koehn 2010). The phrase-based SMT approach offers numerous advantages as a phrase based translation captures word context and local reordering.

## 2.5 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) is a sequence-to-sequence approach (Sutskever et al. 2014) based on encoder-decoder architectures with attention (Bahdanau et al. 2015) or self attention encoder (Vaswani et al.; Wang et al. 2017; 2019a). Given a source sentence  $\mathbf{x}=x_1, x_2, x_3, \dots$  and target sentence  $\mathbf{y}=y_1, y_2, y_3, \dots$ , the training objective for NMT is to maximize the log-likelihood  $\mathcal{L}$  with respect to  $\theta$ :

$$\mathcal{L}_\theta = \sum_{(\mathbf{x}, \mathbf{y}) \in C} \log p(\mathbf{y}|\mathbf{x}; \theta) \quad (2.2)$$

$\theta$  represents set of neural net parameters. The decoder produces one target word at a time by computing the probability

$$p(\mathbf{y}|\mathbf{x}; \theta) = \prod_{j=1}^m p(y_j | y_{<j}, \mathbf{x}; \theta) \quad (2.3)$$

Where  $m$  is the number of words in  $\mathbf{y}$ ,  $y_j$  is the current generated word, and  $y_{<j}$  are the previously generated words. At inference time, beam search is typically used to find the translation that maximises the above probability. Most of NMT models follows the *Embedding*  $\rightarrow$  *Encoder*  $\rightarrow$  *Attention*  $\rightarrow$  *Decoder* framework. More details about attention mechanism can be found in Bahdanau et al. (2015) work.

Unlike traditional SMT, NMT employs a single and massive neural network to model the full translation process. It has the following advantages. First, it captures meaningful syntactic and semantic regularities by the use of a distributed representation of words. Second, it is not necessary to explicitly construct features to capture translation regularities, which is laborious in SMT. Instead, NMT is able to learn representations directly from the training data. Third, NMT captures long-distance reordering, which is a big obstacle for SMT.

### 2.5.1 Multilingual Neural Machine Translation

In recent years, NMT has been shown to be more successful than SMT, thus rapidly becoming the state-of-the-art in MT (Firat et al. 2017). Still, NMT demands a significant amount of parallel data to learn a mapping between languages. To overcome data sparsity, multilingual NMT has attracted attention due to its ability to learn cross-lingual transfer via parameter distribution across multiple languages. This has become an advantages in the case of closely

related languages and under-resourced languages (Lakew et al. 2018).

The multilingual NMT architecture is indistinguishable from the NMT architecture with an optional extension of explicit connections between encoder and decoder layers. One of the simpler and effective multilingual NMT systems was developed by Ha et al. (2016) and Johnson et al. (2017) through introducing a target-forcing token in the input. To have the option to utilise multilingual data within a single system, they proposed a modification to the input data by introducing artificial tokens at the beginning of the input sentence to indicate the source and the target language the model should translate from and to. This enforces the decoder to translate into a specific target language.

An example of a sentence in English to be translated into Tamil would be:

<en> <ta> Translate into Tamil

In a multilingual NMT system, all parameters are inherently shared by all the language pairs being studied. This enables the model to generalise across different language pairs during training. It has been shown that when language pairs with minimal available data and language pairs with plentiful data are blended into a single model, translation quality on the under-resourced language pair is substantially improved (Aharoni et al. 2019).

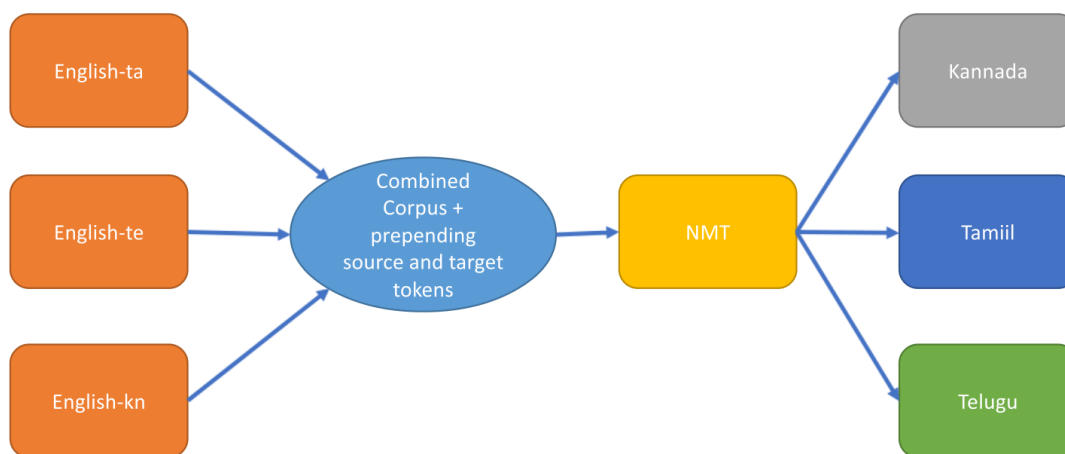


Figure 2.2 – Examples of the multilingual NMT.

## 2.6 Bilingual Lexicon Induction (BLI)

State-of-the-art approaches to bilingual lexicon induction (BLI) use monolingual (Haghighi et al. 2008) or comparable corpora (Fung; Tamura et al. 1995; 2012) to identify pairs of translated words with or without a seed dictionary (Vulić and Korhonen 2016). The induced translation words can improve MT systems (Golan et al. 1988) 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 is based on the 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; Artetxe et al. 2015; 2016) learn a linear mapping  $W$  between the source and target space such that

$$W^* = \operatorname{argmin}_W \sum_i \sum_j D_{ij} \|X_{i*} W - Z_{j*}\|^2 \quad (2.4)$$

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  $V_t$  respectively. Also let  $D \in \{0, 1\}^{|V_s| \times |V_t|}$  be 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.

$$W^* = \operatorname{argmax}_W \operatorname{Tr}(XWZ^T D^T) \quad (2.5)$$

where  $\operatorname{Tr}(\cdot)$  is the trace operator, the sum of all diagonal entries. The optimal solution to this equation is  $W^* = UV^T$ , where  $X^T D Z = U \Sigma V^T$  is the singular value decomposition of  $X^T D Z$ . To get this required seed dictionary, Artetxe et al. (2018a) came up with an iterative, self-learning framework which uses numerals as seed dictionary for the first time to determine  $W$ , uses it to calculate  $D$  and then from next iteration onwards it appends  $D$  to the seed dictionary to learn.

Any method of BLI and dictionary expansion may be used to complement the parallel data used to predict word alignments and scoring phrase tables. The most efficient way to incorporate the lexicon induction output into the SMT pipeline would be to induce translations for rare words and out-of-vocabulary. That is, if there is no translation in the phrase table for a term in our test set, we could induce one for it. Another method is to use BLI to retrieve parallel fragments within the document pair, then supplement existing parallel data with the new sentences to overcome the data sparsity (Irvine and Callison-Burch 2017). In SMT, translation adequacy and fluency can be improved by data augmentation of large amounts of monolingual data (Koehn and Knight 2000a). It has been shown that the injection of under-resourced words into high resource closely related language sentences through an induced bilingual dictionary improves the translation results in two-step pivot translation (Xia et al. 2019).

### 2.7 Automatic Evaluation

Most automatic metrics for MT evaluation are based on an estimation of similarity between system translation and reference translations. In this thesis, we use BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005) and chrF (Popović 2015) metrics for the translation evaluation. BLEU is an automatic evaluation technique which is a geometric mean of  $n$ -gram precision. It is language-independent, fast, and shows a good correlation with human judgment. It is extensively used for various MT evaluations.

The METEOR metric was designed to address the drawbacks of BLEU, and it combines precision and recall. The METEOR metric scores MT system output by aligning them to one or more reference translation. Alignments are based on not only exact word, but also stem, synonym, and paraphrase match between words and phrases. Since word matches between translation and reference include semantic equivalents, it also considers word inflection variation and synonyms. We also used the chrF metric to study system output at the character level, which uses F-score based on character  $n$ -grams. It is language independent and also tokenization independent.

We report BLEU scaled to 0-100 as usual in most research papers although BLEU has initially been defined as 0-1 by Papineni et al. (2002). The higher the BLEU values, the better the translation. METEOR and chrF also follow the BLEU, the higher the values, the better the translation. We also used precision at different levels for the wordnet creation task and accuracy for bilingual lexicon induction.

### 2.8 Summary

In this chapter, we explained the fundamental concepts of our research as essential prerequisites of the thesis. First, we addressed the problem of under-resourced languages. We described the problem and defined some criteria of different orthographic properties influencing MT outputs. In section 2.3, we summarised some of the most important orthographic features addressed to MT. Then we explained the language family under study, Dravidian languages. We also gave a brief overview of SMT, NMT and BLI models. Finally, we end the chapter by discussing the evaluation metrics.

## 3 Literature Review

MT systems are based on Rule-Based Machine Translation (RBMT) or Corpus-Based Machine Translation (CBMT). RBMT systems (Kaji 1988) are based on linguistic knowledge which are encoded by experts. On the other hand, CBMT (Dauphin and Lux; Carl 1996; 2000) depends on a large number of aligned sentences such as SMT (Kondrak et al.; Setiawan et al.; Koehn 2003; 2005; 2005) and NMT (Sutskever et al.; Cho et al. 2014; 2014). Unlike RBMT systems, which require the expertise of linguists to write down the rules for the language, CBMT-based systems rely on examples in the form of sentence aligned parallel corpora. CBMT systems such as SMT and NMT have alleviated the burden of writing down the rules which are not feasible for all languages since human languages are more dynamic in nature.

However, CBMT systems suffer from the lack of parallel corpora for under-resourced languages to train MT systems. A number of methods have been proposed to address the non-availability of parallel corpora for under-resourced languages, such as pivot-based approaches (Wu and Wang 2007), zero-shot translation (Johnson et al. 2017) and unsupervised methods (Artetxe et al. 2019). A large array of techniques have been applied to overcome the data sparsity problem in MT, and virtually all of them are based on the field of transfer learning from high-resource languages. Other techniques are based on lexical and semantic similarities of closely-related languages which are more relevant to this thesis on utilising orthographic information to improve MT of under-resourced languages.

Since our research questions are deeply related to orthographic information, under-resourced language and MT, we began with an introduction to classical types of MT. We review orthographic information in SMT (see Section 3.1), orthographic information in NMT along with multilingual NMT (see Section 3.2), and orthographic information in BLI (see Section 3.3).

### 3.1 Orthographic Information in SMT

The two core methodologies used in the development of MT systems were RBMT and SMT, each come with their own share of advantages and disadvantages. In the initial stages, RBMTs

were the first commercial systems to be developed. These systems were based on linguistic rules and have proved to be more feasible for resource-poor languages with little or no data. It is also relatively simple to carry out error analysis and work on improving the results. Moreover, these systems require very little computational resources.

On the contrary, SMT systems need a large amount of data, but no linguistic theories. Data is especially difficult to collect for the morphologically rich under-resourced languages such as Irish, Malayalam, and Tamil. SMT learns from data and requires less human effort in terms of creating linguistics rules. Even though SMT has lots of advantages over rule-based, it also has some disadvantages. Its is very difficult to conduct error analysis with SMT and data sparsity is another disadvantage faced by SMT (Costa-Jussa et al. 2012). It suffers from out-of-vocabulary problems which are very frequently due to orthographic inconsistencies. To evade this problem, orthographic normalisation was proposed to improve the quality of SMT by sparsity reduction (El Kholy and Habash 2012).

### 3.1.1 Spelling and Typographical Errors

The impact of spelling and typographical errors in SMT has been studied extensively (Bertoldi et al.; Formiga and Fonollosa 2008; 2012). Dealing with random, non-word error or real-word error can be done in many ways; one such method is the use of a character-level translator, which provides various spelling alternatives. Typographical errors such as substitution, insertion, deletion, transposition, run-on, and split can be addressed with edit-distance under a noisy channel model paradigm (Brill and Moore; Toutanova and Moore 2000; 2002). Error recovery is often performed to correct spelling alternative of input before the translation process.

### 3.1.2 True-casing and Capitalization, Tokenization and Detokenization

Most SMT systems accept pre-processed inputs, where the pre-processing consists of tokenising, true-casing, and normalising punctuation. Moses (Koehn et al. 2007) is a toolkit for SMT, which has pre-processing tools for most languages based on hand-crafted rules. Improvement has been achieved for recasing and tokenisation processes (Nakov 2008). For a language which does not use Roman characters, linguistically-motivated tokenisation has shown to improve the results on SMT (Oudah et al. 2019). Byte Pair Encoding (BPE) avoids out-of-vocabulary issues by representing more frequent sub-word as atomic units Sennrich et al. (2016a). A joint BPE model based on the lexical similarity between Czech and Polish identified cognate vocabulary of sub-words. This is based on the orthographic correspondences from which words in both languages can be composed (Chen and Avgustinova 2019).

### 3.1.3 Transliteration (Cognate)

As previously explained, closely-related languages share the same features; the similarities between the language would be of much help to study the cognates of two languages. Several methods have been obtained to manipulate the features of resource-rich languages in order to improve SMT for resource-poor languages. Manipulation of the cognates to obtain transliteration is one of the methods to improve SMT for resource-poor languages.

Language similarities and regularities in morphology and spelling variation motivate the use of character-level transliteration models. However, in order to avoid the character mapping differences in various contexts, [Nakov and Tiedemann \(2012\)](#) transformed the input to a sequence of character n-grams. A sequence of character n-grams increases the vocabulary as well as also make the standard alignment models and their lexical translation parameters more expressive.

For the languages which use the same or similar scripts, approximate string matching approaches, like Levenshtein distance ([Levenshtein 1966](#)) and longest common subsequence ratio (LCSR) ([Melamed 1999](#)) are used to find cognates. For the languages which use different scripts, transliteration is the first step and follow the above approach. A number of studies have used statistical and deep learning methods along with orthographic information ([Ciobanu and Dinu; Mulloni and Pekar 2014; 2006](#)) to find the cognates. In reference to the previous section we know that cognates can be used for mutual translation between two languages if they share similar properties, so it is essential to know the cognateness between the two languages of a given text. The word "cognateness" means how much two pieces of text are related in terms of cognates. These cognates are useful to improve the alignment when the scoring function of the length-based alignment function is very low then it passes to the second method, a cognate alignment function for getting a proper alignment result ([Simard et al. 1993](#)).

One of the applications of cognates before applying MT is parallel corpora alignment. A study of using cognates to align sentences for parallel corpora was done by [Simard et al. \(1993\)](#). Character level methods to align sentences ([Church 1993](#)) are based on a cognate approach ([Simard et al. 1993](#)).

As early as [Bemova et al. \(1988\)](#), researchers have looked into translation between closely-related languages such as from Czech-Russian RUSLAN and Czech-Slovak CESILKO ([Hajic 2000](#)) using syntactic rules and lexicons. The closeness of the related languages makes it possible to obtain an excellent translation by means of straightforward methods. However, both systems were rule-based approaches and bottlenecks included complexities associated with using a word-for-word dictionary translation approach. Nakov and Ng ([Nakov and Ng 2009](#)) proposed a method to use resource-rich closely-related languages to improve the SMT of under-resourced languages by merging parallel corpora and combining phrase tables. The authors developed a transliteration system trained on automatically-extracted likely cognates for Portuguese into Spanish using systematic spelling variation.

Popović and Ljubešić (2014) created an MT system between closely-related languages for the Slavic language family. Language-related issues between Croatian, Serbian and Slovenian are explained by Popović et al. (2016). Serbian is digraphic (uses both Cyrillic and Latin Script), the other two are written using only the Latin script. For the Serbian language transliteration without loss of information is possible from Latin to Cyrillic script because there is a one-to-one correspondence between the characters.

In 2013, Beinborn et al. (2013) used an SMT approach as the base method to produce cognates. Instead of translating the phrase, they tried to transform a character sequence from one language to another. They have used words instead of sentences and characters instead of words in the transformation process. The combination of the phrase table with transformation probabilities, language model probabilities, selects the best combination of sequence. Thus the process includes the surrounding context and produces cognates (Beinborn et al. 2013). It has been demonstrated that the use of cognates improves the translation quality (Kondrak et al. 2003).

### 3.1.4 Code-Switching

An SMT system with a code-switched parallel corpus was studied by Menacer et al. (2019) and Fadaee and Monz (2018) for the Arabic-English language pair. The authors have manually translated or used back-translation methods to translate foreign words. The identification of the language of the word is based on the orthography. For English-Hindi, Dhar et al. (2018) manually translated the code-switched component and shown improvements. MT of social media was studied by Rijhwani et al. (2016) where they tackle the code-mixing for Hindi-English and Spanish-English. The same approach translated the main language of the sentence using the Bing Translate API (Niu et al. 2018).

Back transliteration from one script to native script in code-mixed data is one of the challenging tasks to be performed. Riyadh and Kondrak (2019) adopted three different methods to back transliterate Romanised Hindi-Bangla code-mixed data to Hindi and Bangla script. They have used Sequitur, a generative joint n-gram transducer, DTLM, a discriminate string transducer and OpenNMT<sup>1</sup> a NMT toolkit. Along with these three approaches, they have leveraged target word lists, character language models, as well as synthetic training data, whenever possible, in order to support transliteration. In the end, these transliterations are provided to a sequence prediction module for further processing.

### 3.1.5 Pivot Translation

Pivot translation is a translation from a source language to a target language through an intermediate language which is called a pivot language. Usually, pivot language translation has large source-pivot and pivot-target parallel corpora (Cohn and Lapata; Wu and Wang 2007;

---

<sup>1</sup><https://opennmt.net/>

2009). There are different levels of pivot translation. The first one is the triangulation method where the corresponding translation probabilities and lexical weights in the source-pivot and pivot-target translation are multiplied. In the second method, the sentences are translated to the pivot language using the source-pivot translation system then pivoted to target language using a pivot-target translation system (Utiyama and Isahara 2007). Finally, using the source-target MT system to create more data and adding it back to the source-target model, which is called back-translation (Sennrich et al.; Edunov et al. 2016a; 2018). Back translation is simple and easy to achieve without modifying the architecture of the MT models. Back-translation has been studied in both SMT (Tiedemann et al.; Ahmadnia et al. 2016; 2017) and NMT (Graça et al.; Kim et al. 2019; 2019).

The pivot translation method could also be used to improve MT systems for under-resourced languages. One popular way is training SMT systems using a source-pivot or pivot-target language pair using sub-words where the pivot language is related to source or target or both. The sub-words units consisted of orthographic syllable and byte-pair-encoded unit. The orthographic unit is a linguistically motivated unit which occurs in a sequence of one or more consonants followed by a vowel. Unlike orthographic units, BPE (Byte Pair Encoded Unit) (Sennrich et al. 2016a) is motivated by statistical properties of the text. It represents stable and frequent character sequences in the texts. As orthographic syllable and BPE are variable-length units and the vocabularies used are much smaller than morpheme and word-level model, the problem of data sparsity does not occur but provides an appropriate context for translation between closely-related languages (Kunchukuttan et al. 2017).

## 3.2 Orthographic Information in NMT

In recent years, NMT has improved translation performance, which led to a boom in NMT research. The most popular neural architectures for NMT are based on the encoder-decoder (Bahdanau et al. 2015) architecture and the use of attention or self-attention based mechanism (Luong et al.; Vaswani et al. 2015; 2017). Multilingual NMT created with or without multiway corpora has been studied for the potential for translation between two languages without any direct parallel corpus. Zero-shot translation is translation using multilingual data to create a translation for languages, which have no direct parallel corpora to train independently. Multilingual NMT with only monolingual corpora was studied by Sen et al. (2019) and Wang et al. (2019b). In Ha et al. (2016) and Johnson et al. (2017), the authors have demonstrated that multilingual NMT improves translation quality. For this, they created a multilingual NMT without changing the architecture by introducing special tokens at the beginning of the source sentence indicating the source language and target language.

### 3.2.1 Spelling and Typographical Errors

Spelling errors are amplified in under-resourced setting due to the potentially infinite possible misspellings and lead to a large number of out-of-vocabulary words. Additionally, under-

resourced morphological rich languages have morphological variation, which causes orthographic errors while using character level MT. A shared task was organised by [Li et al. \(2019\)](#) to deal with orthographic variation, grammatical errors and informal languages from the noisy social media text. Data cleaning was used along with suitable corpora to handle spelling errors. [Belinkov and Bisk \(2018\)](#) investigated noise in NMT, focusing on kinds of orthographic errors. Parallel corpora were cleaned before submitting to NMT to reduce the spelling and typographical errors.

NMT with word embedding lookup ignores the orthographic representation of the words such as the presence of stems, prefixes, suffixes and another kind of affixes. To overcome these drawbacks, character-based word embedding was proposed by [Kim et al. \(2016\)](#). Character-based NMT ([Costa-jussà and Fonollosa; Yang et al. 2016; 2016](#)) developed to counteract the disadvantages of languages which do not have explicit word segmentation. This enhances the relationship between the orthography of a word and its meaning in the translation system. For spelling mistake data for under-resourced languages, the quality of word-based translation drops severely, because every non-canonical form of the word cannot be represented. Character-level model overcomes the spelling and typographical error without much effort.

### 3.2.2 True-casing and Capitalization, Tokenization and Detokenization

Although NMT can be trained on end-to-end translations, many NMT systems are still language-specific and require language-dependent pre-processing, such as used in SMT. Moses ([Koehn et al. 2007](#)) a toolkit for SMT has pre-processing tools for most languages which are based on hand-crafted rules. In fact, these are mainly available for European languages. For Asian languages which do not use spaces between words, a segmenter is required for each language independently before feeding into NMT to indicate a word segment. This becomes a problem when we train Multilingual NMT ([Johnson et al. 2017](#)).

A solution for the open vocabulary problems in NMT is to break up the rare words into subword units ([Chitnis and DeNero; Ding et al. 2015; 2019](#)) which has been shown to deal with multiple script languages ambiguities ([Schuster and Nakajima; Wu et al. 2012; 2016](#)). A simple and language-independent tokeniser was introduced for NMT and Multilingual NMT by [Kudo and Richardson \(2018\)](#); it is based on two subword segmentation algorithms, byte-pair encoding (BPE) ([Sennrich et al. 2016a](#)) and a unigram language model ([Kudo 2018](#)). This system also normalises semantically equivalent Unicode characters into canonical forms. Subword segmentation and true-casing models will be rebuilt whenever the training data changes. The pre-processing tools introduced by OpenNMT normalises characters and separates punctuation from words, and it can be used for any language and any orthography ([Klein et al. 2017](#)).

Character-level NMT systems work at the character level to grasp orthographic similarity between the languages. They were developed to overcome the issue of limited parallel corpora and resolve the out-of-vocabulary problem for under-resourced languages. For Hindi-

Bhojpuri, where Bhojpuri is closely-related to Hindi, Bhojpuri is considered as an under-resourced language, and it has a big overlap of word with the high-resource language Hindi (Jha et al. 2019). To solve the out-of-vocabulary problem the transduction of Hindi words to Bhojpuri words was adapted from NMT models by training on Hindi-Bhojpuri cognate pairs. It was a two-level system: first, the Hindi-Bhojpuri system was developed to translate the sentence; then the out-of-vocabulary words were transduced.

### 3.2.3 Transliteration (Cognate)

Transliteration emerged to deal with proper nouns and technical terms that are translated with preserved pronunciation. Transliteration can also be used to improve MT between closely-related languages, which use different scripts since closely-related languages have orthographic and phonological similarities between them.

MT often developed between closely-related languages or through a pivot language (like English) (Bhattacharyya et al. 2016). Translation between closely-related languages or dialects is either a simple transliteration from one language to another language or a post-processing step. Transliterating cognates has been shown to improve MT results since closely-related languages share linguistic features. To translate from English to Finnish and Estonian, where the words have similar orthography, Grönroos et al. (2018) used Cognate Morfessor, a multi-lingual variant of Morfessor which learns to model cognates pairs based on the unweighted Levenshtein distance (Levenshtein 1966). The ideas are to improve the consistency of morphological segmentation of words that have similar orthography, which shows improvement in the translation quality for the resource-poor Estonian language.

Cherry and Suzuki (2009) use transliteration as a method to handle out-of-vocabulary (OOV) problems. To remove the script barrier, Bhat et al. (2016) created machine transliteration models for the common orthographic representation of Hindi and Urdu text. The authors have transliterated text in both directions between Devanagari script (used to write the Hindi language) and Perso-Arabic script (used to write the Urdu language). The authors have demonstrated that a dependency parser trained on augmented resources performs better than individual resources. The authors have shown that there was a significant improvement in BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2002) score and have shown that the problem of data sparsity is reduced.

Recent work by Kunchukuttan et al. (2018) has explored orthographic similarity for transliteration. In their work, they have used related languages which share similar writing systems and phonetic properties such as Indo-Aryan languages. They have shown that multilingual transliteration leveraging similar orthography outperforms bilingual transliteration in different scenarios. Multiway NMT was created for Czech and Polish with Czech IPA transcription and Polish transcription to a 3-way parallel text together to take advantage of the phonology of the closely-related languages (Chen and Avgustinova 2019). Orthographic correspondence rules were used as a replacement list for translation between closely-related Czech-Polish

with added an back-translated corpus (Chen and Avgustinova 2019). Dialect translation was studied by Baniata et al. (2018). To translate Arabic dialects to modern standard Arabic, they used multitask learning which shares one decoder for standard Arabic, while every source has a separate encoder. This is due to the non-standard orthography in the Arabic dialects. The experiments showed that for the under-resourced Arabic dialects, it improved the results.

MT of named entities is a significant issue due to linguistic and algorithmic challenges found in between languages. The quality of MT of named entities, including the technical terms, was improved with the help of developing lexicons using orthographic information. The lexicon integration to NMT was studied for the Japanese and Chinese MT (Halpern 2018). They deal with the orthographic variation of named entities of Japanese using large scale lexicons. For English-to-Japanese, English-to-Bulgarian, and English-to-Romanian Ugawa et al. (2018) proposed a model that encodes the input word based on its NE tag at each time step. This helps to improve the BLEU score for MT results.

### 3.2.4 Code-Switching

A significant part of corpora for under-resourced languages in this thesis comes from movie subtitles and technical documents, which makes it even more prone to code-mixing. Most of these corpora are movie speeches (Birch et al. 2019) transcribed to text, and they differ from that in other written genres: the vocabulary is informal, non-linguistic sounds like *ah*, and mixture of scripts, English and native languages (Tiedemann 2008). Data augmentation (Fadaee et al.; Li and Specia 2017; 2019) and changing the foreign to native words using dictionaries or other methods have been studied. Song et al. (2019) studied the data augmentation method, making code-switched training data by replacing source phrases with their target translation. Character-based NMT (Costa-jussà and Fonollosa; Yang et al.; Lee et al. 2016; 2016; 2017) can naturally handle intra-sentence codeswitching as a result of a many-to-one translation task.

## 3.3 Orthographic Information in Bilingual Lexicon Induction

Building parallel corpora for under-resourced languages is time-consuming and expensive. As a result parallel corpora for under-resourced languages are limited or unavailable for some of the languages. With limited parallel corpora, supervised SMT and NMT cannot achieve the desired quality translations. However, monolingual corpora can be collected from various sources on the Internet, and are much easier to obtain than parallel corpora. Recent research has created a MT system using only monolingual corpora (Koehn and Knight; Ravi and Knight; Dou et al. 2000b; 2011; 2014) by an unsupervised method to remove the dependency of sentence aligned parallel corpora. These systems are based on both SMT (Klementiev et al.; Artetxe et al. 2012; 2018b) and NMT (Artetxe et al. 2018c).

Bilingual lexicon induction is a task of creating word translation from monolingual corpora in

two languages (Turcato; Rosner and Sultana 1998; 2014). One way to induce a bilingual lexicon is using orthographic similarity, based on the assumptions that words that are spelt similarly are sometimes good translations and maybe cognates as they have similar orthography due to historical reasons. A generative model for inducing a bilingual lexicon from monolingual corpora by exploiting orthographic and contextual similarities of words in two different languages was proposed by Haghghi et al. (2008). Many methods, based on edit-distance and orthographic similarity are proposed for using linguistic features for word alignments using supervised and unsupervised methods (Dyer et al. 2011). Riley and Gildea (2018) proposed a method to utilise orthographic information in word-embeddings based bilingual lexicon induction. The authors used the two languages' alphabets to extend the word embeddings and modify the similarity score functions of previous word-embedding methods to include the orthographic similarity measure. Bilingual lexicons are shown to improve MT in both RBMT (Turcato 1998) and CBMT (Chu et al. 2014).

In work by Bloodgood and Strauss (2017), the authors translated lexicons for a heavily code-switched text of historically unwritten colloquial words via loanwords using expert knowledge with language information. Their method is to take word pronunciation (IPA) from a donor language and convert them into the borrowing language. This shows improvements in BLEU score for induction of a Moroccan Darija-English translation lexicon bridging via French loan words.

### 3.4 Related Work on Wordnet Creation

In this section, we move on to related work on wordnet creation since it is one of our main contributions. IndoWordNet covers the official Indian languages, from the major three families: Indo-Aryan, Dravidian and Sino-Tibetan languages. In general, Indian languages are rich in morphology, and each of the three language families has different morphology structure. It was compiled for eighteen out of the twenty-two official languages and made publicly available.<sup>2</sup> Similarly to EuroWordNet, the central language is Hindi, which is then linked to English. The IndoWordNet entries are updated frequently. For the Tamil language, Rajendran et al. (2002) proposed a design template for the Tamil wordnet. In their further work (Rajendran et al. 2010), they emphasise the need for an independent wordnet for the Dravidian languages, based on EuroWordNet. This is due to the observation that the morphology and lexical concepts of these languages are different compared to other Indian languages. The authors have combined the Tamil wordnet and wordnets in other Dravidian languages to form the IndoWordNet.

Mohanty et al. (2017) built SentiWordNet for the Odia language, which is one of the official languages of India. Being an under-resourced language, Odia lacks a proper MT system to translate the vocabulary of the available resource from English. The authors have created SentiWordNet for Odia using the resources of other Indian languages and the IndoWordNet. The IndoWordNet structure does not map directly to the SentiWordNet, so instead synsets

---

<sup>2</sup><http://www.cfilt.iitb.ac.in/indowordnet/index.jsp>

are matched. The authors used these for translation from a source lexicon to a target lexicon. Similarly, the work by [Horváth et al. \(2016\)](#) focuses on the semi-automatic construction of wordnet for the Mansi language, which is spoken by Mansi people in Russia, an endangered under-resourced language with a low number of native speakers. The authors have used the Hungarian wordnet as a starting point. With the help of a Hungarian-Mansi dictionary, which was used to create possible translations between the languages, the Mansi wordnet was continuously expanded.

Previous works made lots of manual effort to create wordnet-like resources, which was funded by public research for an extended period of time. However, IndoWordNet is not complete and biased towards Hindi, because the authors created a Hindi-Tamil bilingual dictionary, rather than a wordnet. As explained in [Rajendran et al. \(2010\)](#), the morphology and lexical concepts of Dravidian languages are different from Hindi, which illustrates that the IndoWordNet may not be the most suitable resource to represent a wordnet for the targeted Dravidian languages.

To evaluate and improve the wordnets for the targeted Dravidian languages, we follow the approach of [Arcan et al. \(2016b\)](#), which uses the existing translations of wordnets in other languages to identify contextual information for wordnet senses from a large set of generic parallel corpora. We use this contextual information to improve the translation quality of WordNet senses. We show that our approach can help overcome the drawbacks of simple translations of words without context.

### 3.5 Summary

In this chapter, we provided detailed related work. We first reviewed the orthographic information in SMT systems. We then discussed the orthographic information in NMT and multilingual NMT. We also gave information about bilingual lexicon induction and literature related to our thesis. Although our brief survey shows that utilising orthographic information in SMT, NMT, and BLI has been successfully applied in different settings, our approach differs from past works in several aspects: i) introducing a method to remove the code-mixed based on scripts at both sides of the parallel corpus, where the size of the corpus gets reduced, however, this approach improves results, ii) previous approaches studied multilingual NMT and BLI for languages which use a similar script, in this thesis we consider closely-related languages which have dissimilar scripts and, iii) we compare the effect of different scripts in multilingual NMT and show Latin script outperforms IPA and native script.

# Code Mixing **Part III**



## 4 Improving Word Sense Translation by Removing Code-mixing

### 4.1 Introduction

In the previous part II, we have introduced the necessary background to understand MT and literature related to orthographic information in MT. In this chapter, we present a method to deal with code-mixing in a parallel corpus, and we created a method to improve wordnets for under-resourced languages. This is the first step towards utilizing orthographic information in MT, which directly addresses our research question (RQ1).

Code-switching or code-mixing is a phenomenon found among bilingual communities all over the world (Ayeomoni; Yoder et al. 2006; 2017). Code-mixing is the mixing of words, phrases, and sentences from two or more languages within the same sentence or between sentences. In many bilingual or multilingual communities like India, Hong Kong, Malaysia or Singapore, language interaction often happens in which two or more languages are mixed. Furthermore, it increasingly occurs in monolingual cultures due to globalization. Due to the history and popularity of the English language on the Internet, Indian languages are more frequently mixed with English than other native languages (Chanda et al. 2016). While using any language, there is a trend of borrowing from other languages. A borrowed word or loan word from another language initially shows up more frequently in speech, and then continuously in print media like newspaper, and finally, it loses its originality. It becomes part of a language resulting in an inclusion in the dictionary of a language. Borrowed words are therefore added to the lexicon of a new language. Usually, it takes several years for borrowed words to formally become part of a language's dictionary (Patro et al. 2017). Code-mixing is different from borrowing since it keeps the identity of each language. Speakers whose first language uses a non-Latin script write using the Latin script due to the accessibility of modern digital devices keyboard (phonetic typing) which also increases the probability of code-mixing (Das and Gambäck 2013).

Code-mixing can be categorized into inter-sentential and intra-sentential switching (Hamed et al. 2018). Despite the fact that code-switching is considered to be the synonym of code-mixing, the former refers to the alteration of two or more languages in the same sentence (*intra-*

## Chapter 4. Improving Word Sense Translation by Removing Code-mixing

---

*sentential code-mixing*) whereas the latter alludes to the same act between sentences (*inter-sentential code-mixing*) (Barman et al. 2014a). For example:

1. *nee naale evng shopping varunnundo?*

Are you coming for shopping tomorrow evening?

2. *I do not agree. Unakkepadi theriyum?*

I do not agree. How do you know this?

The sentences quoted in italics are a mix of English with Malayalam and Tamil respectively, the languages spoken in the southern region of India. The first one is an example for intra-sentential-code-mixing as the mixing of English words is done inside the same sentence. The other one shows inter-sentential-code-mixing where the speaker alternates language between sentences.

Since under-resourced languages lack corpora, we often depend on the data from voluntary annotators, which often contains code-mixing. A major part of our corpora comes from movie subtitles and technical documents, which makes it even more prone to code-mixing of English in the Dravidian languages. In our corpus, movie speeches are transcribed to text, and they differ from that in other written genres: the vocabulary is informal, non-linguistics sounds like *ah*, and mixing of scripts in case of English and native languages (Tiedemann 2008). Two examples of code-mixing are demonstrated in Figure 4.1.

Source sentence: “இப்போது, நான் அதை loving.”

Transliteration: :Ippōtu, nāṅ atai loving

Target sentence: “Right now, I'm loving it.”

Source sentence: “முன்னிருப்பு GNOME பொருள்”

Transliteration: :Munṅiruppu GNOME poruḷ

Target sentence: “Default GNOME Theme”

Figure 4.1 – Examples of code-mixing in Tamil-English parallel corpus. In the first example the verb *loving* is code-mixed in Tamil. In second example the noun *GNOME* is code-mixed.

As computational activities and the Internet create a wider multilingual and global community, under-resourced languages acquire political as well as an economic interest to develop NLP systems for these languages. In general, creating NLP systems requires an extensive amount of resources and manual effort; however, under-resourced languages lack in both. One such essential resource is a wordnet (Miller; Vossen 1995; 1997), which has been utilized for various purposes in NLP, including word-sense disambiguation, text classification, text summarizing, information retrieval, and MT. Wordnets are lexical resources, which provide a hierarchical structure based on synsets (a set of one or more synonyms) and semantic features of individual words. A classical utilization of wordnet is to decide the similarity between words.

Wordnets can be constructed by either the *merge* or the *expand* approach (Vossen 1997). Princeton WordNet (Miller; Fellbaum 1995; 2010) was manually created by Princeton University covering the vocabulary of English only. Then, based on the Princeton WordNet, wordnets for several languages were created. As an example, EuroWordNet (Vossen 1997) is a multilingual lexical database for several European languages, structured in the same way as Princeton’s WordNet. The MultiWordNet (Pianta et al. 2002) is strictly aligned with Princeton WordNet and allows access senses in Italian, Spanish, Portuguese, Hebrew, Romanian and Latin. Many others have followed for different languages. The IndoWordNet (Bhattacharyya 2010) was compiled for eighteen out of the twenty-two official languages of India and made available for public use. It is based on the *expand* approach like EuroWordNet, but from the Hindi wordnet, which is then linked to English. On the Global WordNet Association website,<sup>1</sup> a comprehensive list of wordnets available for different languages can be found, including IndoWordNet and EuroWordNet.

In this chapter, we seek to answer our first research question:

**RQ1:** Does removing foreign words (code-mix) based on orthography improve word sense translation quality?

To answer RQ1, we describe the effort towards generating and improving wordnets for Dravidian languages such as Tamil, Telugu and Kannada by removing foreign words in the training corpora. Since studies (Federico et al.; Läubli et al.; Green et al. 2012; 2013; 2013) have demonstrated significant efficiency gains when human translators post-edit MT output instead of translating text from scratch, we use the available parallel corpora from multiple sources, like OPUS,<sup>2</sup> to create an MT system to translate the wordnet senses in the Princeton WordNet into the mentioned under-resourced languages. The parallel corpora from OPUS had many code-mixing points in the corpora. The handling of code-mixing by removing them based on the orthography appears to improve the quality of the results outperforming the baseline results of wordnet entries. Thus, we believe that the method presented here is applicable to resource creation of under-resourced languages with minimum effort. Translation tools such as Google Translate,<sup>3</sup> or open-source SMT systems such as Moses (Koehn et al. 2007) trained on generic data are the most common solutions, but they often result in unsatisfactory translations of domain-specific expressions. Therefore, we follow the idea of Arcan et al. (2016b), where the authors automatically identify relevant sentences in English containing the WordNet senses and translate them within the context, which showed translation quality improvement of the targeted entries. The effectiveness of our approach is evaluated by comparing the generated translations with the IndoWordNet entries, automatically and manually, respectively. This chapter reports our first outcomes in improving wordnet for under-resourced Dravidian languages such as Tamil, Telugu and Kannada.

---

<sup>1</sup><http://globalwordnet.org/>

<sup>2</sup><http://opus.lingfil.uu.se/>

<sup>3</sup><http://translate.google.com/>

## Chapter 4. Improving Word Sense Translation by Removing Code-mixing

Publication from Chapter 4:

- Bharathi Raja Chakravarthi, Mihael Arcan and John P. McCrae "Improving Wordnets for Under-Resourced Languages Using Machine Translation", Proceedings of the 9th Global WordNet Conference, (2018).

### 4.2 Our Approach

Our specific aim of this work is to generate and improve wordnets for under-resourced languages without foreign words. The principle approaches for constructing wordnets are the *merge* approach or the *expand* approach. In the *merge* approach, the synsets and relations are built independently and then aligned with WordNet. The drawbacks of the *merge* approach are that it is time-consuming and requires much manual effort to build.

On the contrary, in the *expand* model, wordnet can be created automatically by translating synsets using different strategies, whereby the synsets are built in correspondence with the existing wordnet synsets. For our task, we chose the *expand* approach and automatically translated the Princeton WordNet entries within a disambiguate context to obtain entries for the Dravidian languages. Orthographic information was utilized to remove foreign words from the corpora to improve MT quality.

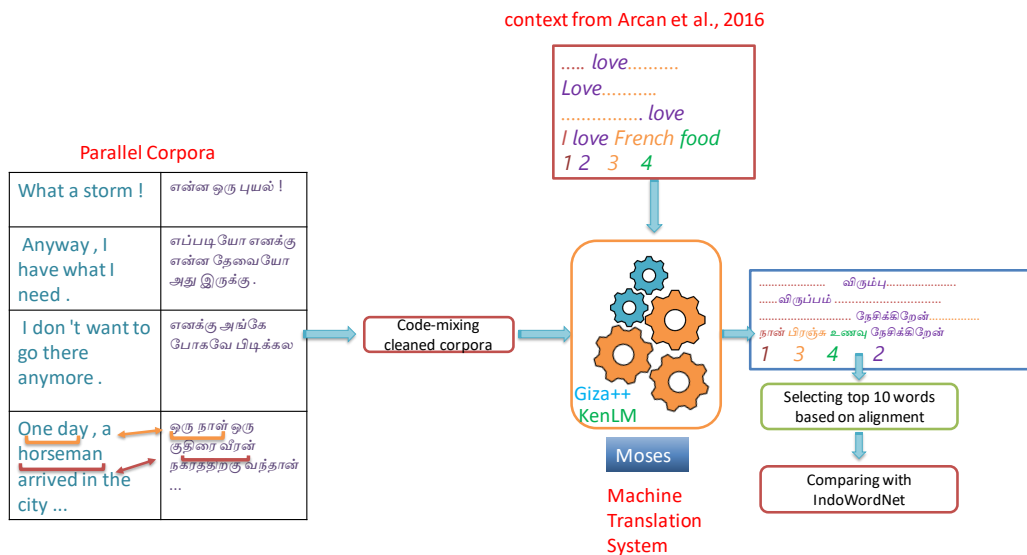


Figure 4.2 – Work flow of creating a wordnet for Dravidian languages by cleaning up code-mixed corpora.

To improve the wordnets for the targeted Dravidian languages, we collect the parallel corpora from publicly available resources, we then remove the code-mixing to create noise-free

corpora. We use SMT to create an MT system for Dravidian languages from English. We then pursue the approach of Arcan et al. (2016b), which utilizes the existing translations of wordnets in different languages to identify contextual information for wordnet senses from a extensive set of generic parallel corpora. We use this contextual information to improve the translation quality of WordNet senses. In the following subsections, we explain our approach in detail.

### 4.2.1 Available Corpora for MT

This section describes the data collection and the pre-processing steps. The English-Tamil parallel corpus, which we used to train our SMT system is collected from various sources and combined into a single parallel corpus. We used the EnTam corpus (Ramasamy et al. 2012), which was pre-processed from raw Web data to become a sentence-aligned corpus. The parallel corpora contain text from the news domain,<sup>4</sup> sentences from the Tamil cinema articles<sup>5</sup> and the Bible.<sup>6</sup> For the news corpus, the authors downloaded web pages that have matching filenames in both English and Tamil. For the cinema corpus, all the English articles had a link to the corresponding Tamil translation. The collection of the Bible corpus followed a similar pattern. We also took the English-Tamil parallel corpora for six Indian languages created with the help of Mechanical Turk for Wikipedia documents (Post et al. 2012). Since the data was created by non-expert translators hired over Mechanical Turk, it is of mixed quality. From the OPUS website, we have collected the Gnome, KDE, Ubuntu and movie subtitles (Tiedemann 2012). We furthermore manually aligned the Tamil text Tirukkural,<sup>7</sup> and combined all the parallel corpora into a single corpus. We first tokenized sentences in English and Tamil and then true-cased only the English side of the parallel corpus, since the Tamil language does not have a casing. Finally, we cleaned up the data by eliminating the sentences whose length is above 80 words.

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
Number of tokens	7,738,432	6,196,245	258,165	226,264	68,197	71,697
Number of unique words	134,486	459,620	18,455	28,140	7,740	15,683
Average word length	4.2	7.0	3.7	4.8	4.5	6.0
Average sentence length	5.2	7.9	4.6	5.6	5.3	6.8
Number of sentences	449,337		44,588		13,543	

Table 4.1 – Statistics of the parallel corpora used to train the translation systems.

To obtain the parallel corpus for Telugu and Kannada, we used the corpora available on the OPUS website. The same pre-processing procedure was followed for Telugu and Kannada language since both languages are close to the Tamil language. Table 4.1 shows the statistics

<sup>4</sup><http://www.wsws.org/>

<sup>5</sup><http://www.cinesouth.com/>

<sup>6</sup><http://biblephone.intercer.net/>

<sup>7</sup><http://www.projectmadurai.org/>

of the parallel corpora for the three language pairs. From this table, we can see that the English-Tamil parallel corpus is much larger than for the other language pairs. On the other hand, the number of sentences for English-Kannada is very small. Once we have obtained the parallel corpus, we created the SMT systems for the English-Tamil, English-Telugu, and English-Kannada language pairs.

We define the following set of data:

- Development set: 2,000 randomly selected sentences from the parallel corpus are used as a development set to measure the system performance of the phrase-based translation model.
- Test set: A blind set of 1,000 sentences randomly chosen from a parallel corpus are used to test the system. There is no overlap between these sets of data.
- Training set: A larger size parallel corpus that is used to train the phrase-based translation model. It is the remaining corpus after the development and test set are extracted.

In this work, we focus on three languages from the Dravidian family, namely, Tamil, Telugu, and Kannada. This is mainly due to the availability of parallel corpora, and we believe that this method can be extended to other under-resourced languages without much effort.

### 4.2.2 Resource Scarceness

There are few resources, which can be used to create a wordnet for under-resourced languages automatically. One way to cross the language barrier is with the help of MT. As with any machine learning method, SMT tends to improve translation quality when using a large amount of training data. That is, if the training method sees a specific word or phrase multiple times during training, it is more likely to learn the correct translation. SMT suffer due to the scarcity of parallel corpora, Dravidian word order and the morphological complexity of the language. For the Dravidian languages when translating from or to English, the translation models suffer because of syntactic differences while the morphological differences contribute to data sparsity. In contrast, small corpora used for training lead to incomplete word coverage, which may cause out-of-vocabulary (OOV) issues.

Besides the resource scarceness, another issue observed with the corpus for Dravidian languages was code-mixing contents in the data. Code-mixing is an act of alternating between elements of two or more languages, which is prevalent in multilingual countries ([Barman et al. 2014b](#)). With English being the most used language in the digital world, people tend to mix English words with their native languages. That might be the case in other languages as well.

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
tok	0.5% (45,847)	1.1% (72,833)	2.8% (7,303)	4.9% (12,818)	3.5% (2,425)	9.0% (6,463)
sent	0.9% (4,100)		3.1% (1,388)		3.4% (468)	

Table 4.2 – Number of sentences (sent) and number of tokens (tok) removed from the original corpus.

### 4.2.3 Code-mixing

A major part of our corpora comes from movie subtitles and technical documents, which makes it even more prone to code-mixing of English in the Dravidian languages. The parallel corpus is initially divided into English script and native script. All of the annotations are done using an automatic process. All words from a language other than the native script of our experimental data are taken out on both sides of the corpus if it occurs in the native language side of the parallel corpus. The sentences are removed from both sides of the target language side does not contain native script words in it. An example is given in Figure 4.3.

#### Baseline

Source sentence:	“இப்போது, நான் அதை loving.”
Transliteration:	:lppōtu, nāṅ atai loving.
Target sentence:	“Right now, I'm loving it.”
English/Tamil	

#### After Code-mixing removed

Source sentence:	“இப்போது, நான் அதை.”
Transliteration:	:lppōtu, nāṅ atai.
Target sentence:	“Right now, I'm it.”
English/Tamil	

Figure 4.3 – Examples of the before and after removal of code-mix from the corpora.

Table 4.2 shows the percentage of code-mixed text removed from the original corpus. The goal of this approach is to investigate whether code-mixing criteria and corresponding training are directly related to the improvement of the translation quality measured with automatic evaluation and manual evaluation. We assumed that code-mixed text can be found by different scripts and did not evaluate the code-mixing written in the native script or Latin script to write the native language as was done by [Das and Gambäck \(2013\)](#).

### 4.2.4 Context Identification

Since the manual translation of wordnets using the extend approach is a very time consuming and expensive process, we apply SMT to translate WordNet gloss entries which were collected

by Arcan et al. (2016b) for context identification. The authors used the Princeton WordNet for English and Open Multilingual Wordnet for the other ten languages, Croatian, Dutch, Finnish, French, Italian, Portuguese, Romanian, Slovene, and Spanish. They have used existing translations of WordNet in other languages to identify contextual information for wordnet senses from a large set of generic parallel corpora.

We use this contextual information sentences from Arcan et al. (2016b) work to translate between English and the target Dravidian languages using SMT. While a domain-unadapted SMT system can only return the most frequent translation when given a term by itself, it has been observed that translation quality of single-word expressions improves when the word is given in a disambiguated context of a sentence (Arcan et al.; Arcan et al. 2016a; 2016b). Therefore existing translations of WordNet senses in languages other than English were used to select the most relevant sentences for wordnet senses from a large set of generic parallel corpora. The goal is to identify sentences that share the same semantic information in respect to the synset of the WordNet entry that we want to translate. To ensure a broad lexical and domain coverage of English sentences, existing parallel corpora for various language pairs were merged into one parallel data set, i.e., Europarl (Koehn 2005), DGT - translation memories generated by the *Directorate-General for Translation* (Steinberger et al. 2014), MultiUN corpus (Eisele and Chen 2010), EMEA, KDE4, OpenOffice (Tiedemann 2009), OpenSubtitles2012 (Tiedemann 2012). Similarly, wordnets in a variety of languages, provided by the Open Multilingual Wordnet web page,<sup>8</sup> were used.

### 4.2.5 Training MT parameters

In the following section, we take a baseline noisy parallel corpus and non-code mixed parallel corpus, that has been aligned at the sentence level. To obtain the translations, we use the Moses SMT toolkit with a baseline setup with 5-gram language model created using the training data by KenLM (Heafield 2011). The baseline and non-code mixing SMT systems were built for three language pairs, English-Tamil, English-Telugu, and English-Kannada. The test set mentioned in Section 4.2.1 was used to evaluate our systems. From Table 4.1 and Table 4.3 we can see that there is no significant reduction in BLEU even after removal of noisy sentence.

	Original	Non-Code mixing
English→Tamil	20.29	20.61
English→Telugu	28.81	28.25
English→Kannada	14.64	14.45

Table 4.3 – Automatic translation evaluation of the of 1,000 randomly selected sentences in terms of the BLEU metric.

---

<sup>8</sup><http://compling.hss.ntu.edu.sg/omw/>

## 4.3 Results

The most reliable method to evaluate a wordnet is a manual evaluation, but a manual evaluation of the whole wordnet is time-consuming and very expensive. Therefore, we did the automatic evaluation of our translations and measured the precision. In order to determine the correctness of our work, we have furthermore randomly taken 50 wordnet entries for manual evaluation on these entries.

### 4.3.1 Automatic Evaluation

In this section, we have compared our result to the IndoWordNet. Once the translation step of disambiguated context containing the target entries was finished, we then used the word alignment information to extract the translation of the WordNet entry. Since several disambiguated sentences per WordNet entry were used, we took the translations for each context and then combined the results to count the most frequent one.

The goal of this work is to aid the human annotator in speeding up the process of wordnet creation for under-resourced languages. Precision at different levels is calculated by comparing it with IndoWordNet. For precision at 10, the top 10 entries were compared to the IndoWordNet by checking whether at least one of them matched exactly, similarly for precision at different levels. We took precision at 10, precision at 5, precision at 2, and precision at 1 for our work. We made this comparison for all the three languages, i.e. Tamil, Telugu, and Kannada. We removed the code-mixed part of the corpus and created a new translation system, which was used again to translate the same WordNet entries. Table 4.4 shows the result of the automatic evaluation of the translation of the entries into the targeted Dravidian languages. The table shows the precision at the different level of the translations, based on the translation model, generation from the original corpus and non-code mixed corpus. Non-code mixed often outperforms the baseline in terms of precision, whereas the difference is less visible in the Telugu language. This is likely due to the short sentences in the Telugu corpus. These differences in the precision are better in the manual evaluation of Tamil tests with 50 samples. The vast difference between manual and automatics evaluation can be explained in part by different forms.

Table 4.4 shows an example of how our system differs from the baseline SMT system and how it benefits the wordnet translation. This is clear evidence that SMT without code-mixing described above achieves an improvement over the baseline without using any additional training data. However, it has been shown in [Arcan et al. \(2016b\)](#) that better performance on WordNet translation can be achieved if the corpora contained a sufficient amount of parallel sentences. Their translation evaluation based on the BLEU metric on unigrams (similar to precision at 1, P@1), showed a range between 0.55 and 0.70 BLEU points, for the well-resourced languages, like Slovene, Spanish, Croatian and Italian. Restricting the task to a small data set tends to hurt the translation performance, but it can be useful to aid in the creation or improvement of new resources for the under-resourced languages.

		English→Tamil			
		P@10	P@5	P@2	P@1
original corpus		0.120	0.109	0.083	0.065
non-code mixed		0.125	0.115	0.091	0.073
		English→Telugu			
		P@10	P@5	P@2	P@1
original corpus		0.047	0.046	0.038	0.028
non-code mixed		0.047	0.045	0.038	0.027
		English→Kannada			
		P@10	P@5	P@2	P@1
original corpus		0.009	0.010	0.008	0.005
non-code mixed		0.011	0.011	0.009	0.007

Table 4.4 – Results of Automatic evaluation of wordnet with IndoWordNet Precision at different level denoted by P@10 which means Precision at 10.

### 4.3.2 Manual Evaluation

In order to be able to evaluate our method in contrast to stand-alone approaches, we manually evaluated our method in comparison with IndoWordNet entries. To select the sample for manual evaluation, we proceeded as follows: we randomly extracted a sample of 50 wordnet entries from the WordNet. First, each of these 50 wordnet entries was compared to the IndoWordNet for the exact match. Subsequently, regardless of this decision, each of the 50 wordnet entries was evaluated and classified according to its quality. The classification is the following:

- **Agrees with IndoWordNet** Exact match found in IndoWordNet.
- **Inflected form** The root of a word is found with a different inflection, which can make the translation correct but imprecise.
- **Transliteration** The word is transliterated, which can be caused by the unavailability of the translation form in the parallel corpus, since some words are used in transliteration because they are borrowed words.
- **Spelling Variant** Since our data comes from day to day language (spoken form) of Tamil but IndoWordNet is skewed towards the classical sense of language (written form). Our method produces spelling variants which can be caused by spelling difference in spoken and written of the word.
- **Correct, but not in IndoWordNet** IndoWordNet is large, and it covers eighteen languages, but it lacks some wordnet entries for the Dravidian languages. We verified that we had identified the correct sense by referring to the wordnet gloss.

ILI code	Gloss	IWN	Meaning	Translation	Meaning	Comments
14647235-n	any of several compounds containing chlorine and nitrogen; used as an antiseptic in wounds	நைட்ரஜன்	nitrogen	நைதரசன்	nitrogen	Spelling variant
01026095-v	give the name or identifying characteristics of; refer to by name or some other identifying characteristic	பெயரிடு	name, identity	பெயர்	name	Inflected form, different part-of-speech
00461782-n	a game in which balls are rolled at an object or group of objects with the aim of knocking them over or moving them	பந்து	ball	பௌலிங்	bowling	Correct translation, sense missing in IWN
04751305-n	noticeable heterogeneity	பல்வேறு	diverseness, diversity	பல்வேறு	diverseness, diversity	Agrees with IWN
01546111-v	be standing; be upright	தூக்கு	to lift	நிற்க	to stand	correct translation, sense missing in IWN

Figure 4.4 – Examples of the manual evaluation of Tamil wordnet entries in comparison to the IndoWordNet (IWN).

- **Incorrect** This error class can be caused due to inappropriate term or mistranslation.

	Original	Non-Code mixing
Agrees with IWN	18%	20%
Inflected Form	12%	22%
Transliteration	4%	4%
Spelling variant	2%	2%
Correct, but not in IWN	18%	24%
Incorrect	46%	28%

Table 4.5 – Manual Evaluation of wordnet creation for Tamil language compared with IndoWordNet (IWN) at precision at 10 presented in percentage.

The examples in Figure 4.4 list the manual evaluation of Tamil wordnet entries in comparison to the IndoWordNet. Neither the word nor its translation has appeared in the training corpus. Therefore, the SMT system cannot translate the word and chooses to produce the word in English.

We should note that this evaluation was carried out for both the original, uncleaned, corpus as well as the cleaned corpus (non-code mixing). We observed that the cleaned corpus produce better results compared to the original data, which have many code-mixing entries. From Table 4.5, we can see that there is a significant improvement over the inflected form and correct but not found in IndoWordNet categories. This shows that our method can help to improve the wordnet entries for under-resourced languages.

### 4.4 Discussion and Conclusion

While our automatic evaluation results are a little disappointing, and this is perhaps unsurprising in the context of under-resourced languages as there is very little data availability for these languages, our manual evaluation shows that this is far from reality. Evaluating using a resource such as IndoWordNet is always likely to be problematic as the resource is far from complete and does not claim to cover all words in the Dravidian languages studied in this chapter. Moreover, IndoWordNet is overly skewed to the classical words of these languages, but the majority of our parallel corpus consists of day to day conversation texts. Despite the low precision in determining the exact match to the IndoWordNet, our technique yields 48% for precision at 10 in manual evaluation, although the automatic evaluation considering precision at 10 gave only 12%. Our method relies on IndoWordNet for evaluation, but IndoWordNet is biased over one particular language, which is Hindi. The resulting wordnet entries, though noisy, are suitable for aiding wordnet creation for under-resourced languages. The handling of code-mixing in this chapter appears to improve the quality of the proposed translation, outperforming the baseline results of wordnet entries once code-mixed was removed from the data. Thus we believe that the method presented here is still applicable to resource creation of under-resourced languages.

### 4.5 Summary

In this chapter, we show the experiments done to explore RQ1. We describe the effort towards generating and improving wordnets for Dravidian languages such as Tamil, Telugu and Kannada. We used available parallel corpora from multiple sources, such as OPUS to create an MT system to translate the wordnet senses in the Princeton WordNet into the above mentioned under-resourced languages. The parallel corpora from OPUS had many code-mixing points in the corpora. The handling of code-mixing by removing them based on the orthography appears to improve the quality of the results, outperforming the baseline results of wordnet entries. Thus, we believe that the method presented in this work is applicable to resource creation of under-resourced languages with minimum effort. We also introduced a new manual evaluation metric to evaluate the wordnet creation task.

However, MT is an active research area in which technologies are being developed expeditiously. With the availability of supercomputer for MT research community, neural network-based MT became the state-of-the-art in MT task.

# Phonetic Transcription **Part IV**



# 5 Comparison of Different Orthography

## 5.1 Introduction

In Part III, we studied the effect of removal of code-mixing in SMT for word sense translation. In this part, we proceed onward to one more property of orthographic information which is the phonetic transcription to improve NMT for under-resourced languages. Phonetic transcription is a method for writing the language in another script keeping the phonemic units intact. It is extensively used in speech processing research, text-to-speech, and speech database construction. Phonetic transcription into a single script has the advantage of collecting similar words at the phoneme level. In this chapter, we propose to alleviate the problem of different scripts by transcribing the native script into a common representation, i.e. the Latin script or the IPA. An example of phonetic transcription of Tamil is given in Example 5.1.

Native script: உங்களுக்கு வணக்கம்  
Latin script: ungaluku vanakam  
International Phonetic Alphabet : ʊŋgəɻkkə ʋəŋəkkəm  
Translation: Hello to you

Figure 5.1 – Examples of phonetic transcription.

Under-resourced languages are a significant challenge for statistical approaches to MT, and recently it has been shown that the usage of training data from closely-related languages can improve MT quality of these languages (Popović and Ljubešić; Abe et al. 2014; 2018). Closely-related languages refer to languages that share similar lexical and structural properties due to sharing a common ancestor (Popović et al. 2016). Frequently, languages in contact with other language or closely-related languages like the Dravidian, Indo-Aryan, and Slavic family share words from a common root (*cognates*), which are highly semantically and phonologically similar. For example, people from Sweden, Denmark, and Norway can communicate with each other without the prior knowledge of other languages since these languages are closely related

## Chapter 5. Comparison of Different Orthography

(Gooskens 2007). While languages within the same language family share many properties, many under-resourced languages are written in their own native script, which makes taking advantage of these language similarities difficult. Examples of this are shown in Figure 5.2

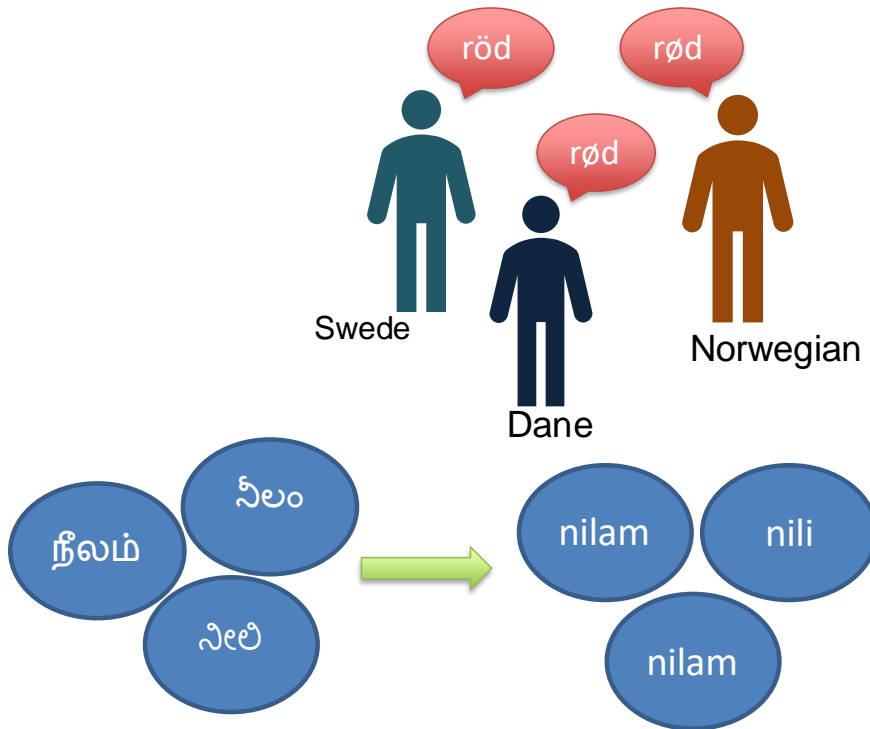


Figure 5.2 – Example of closely related languages

In this chapter, we attempt to investigate the approach of Multilingual NMT (Ha et al. 2016), in particular, the *multi-way* translation model (Firat et al. 2016), where multiple source and target languages are trained simultaneously. This has been shown to improve the quality of the translation. However, in this work, we focus on languages with different scripts, which limits the application of these multi-way models. In order to overcome this, we investigate if converting them into a single script will enable the system to take advantage of the phonetic similarities between these closely-related languages. In this chapter, we seek to answer [RQ2]

**RQ2:** Does phonetic transcription of parallel corpora help to take advantage of resources from closely related language to improve translation quality?

We compare the difference between coarse-grained transliteration into Latin script and fine-grained IPA transliteration. We performed experiments on translation between the language pairs English-Tamil, English-Telugu, and English-Kannada. Our results show improvements in terms of the BLEU, METEOR and chrF scores from transliteration and we find that the transliteration into the Latin script outperforms fine-grained IPA transcription.

This work was published as:

- Bharathi Raja Chakravarthi, Mihael Arcan and John P. McCrae “[Comparison of Different Orthographies for Machine Translation of Under-resourced Dravidian Languages.](#)” In proceedings of the 2nd Conference on Language, Data and Knowledge, Leipzig, Germany (2019).

Our proposed methodology allows the creation of MT systems from under-resourced languages to English and in other direction. Our results, presented in Section 5.3, show that phonetic transcription of parallel corpora increases the MT performance in terms of the BLEU, METEOR and chrF metric (Chakravarthi et al. 2018). Multilingual NMT with closely-related languages improve the score, and we demonstrate that transliteration to Latin script outperforms the more fine-grained IPA.

## 5.2 Our Approach

### 5.2.1 Data

To train an NMT system for the English-Tamil, English-Telugu, and English-Kannada language pairs, we use parallel corpora from OPUS<sup>1</sup> (Tiedemann and Nygaard 2004) as we did in Chapter 4. OPUS includes a large number of translations from the EU, open-source projects, the Web, religious texts and other resources. OPUS also contains translations of technical documentation from the KDE, GNOME, and Ubuntu projects. We took the English-Tamil parallel corpora created with the help of Mechanical Turk for Wikipedia documents (Post et al. 2012), EnTam corpus (Ramasamy et al. 2012) and manually aligned the well-known Tamil text Tirukkural, which contains 2,660 lines. Most multilingual corpora come from the parliament debates and legislation of the EU or multilingual countries, but most non-EU languages lack such resources. For our experiments, we combined all the corpora to form a multilingual corpus and split the corpora into an evaluation set containing 1,000 sentences, a validation set containing 1,000 sentences, and a training set containing the remaining sentences shown in Table 5.1. Following Ha et al. (2016), we indicate the language by prepending two tokens to indicate the desired source and target language.

An example of a sentence in English to be translated into Tamil would be:

```
<en> <ta> Translate into Tamil
```

---

<sup>1</sup><http://opus.nlpl.eu/>

## Chapter 5. Comparison of Different Orthography

	Number of sentences	Tokens-English	Tokens-Dravidian
English-Tamil	2,248,685	44,139,295	34,111,290
English-Telugu	224,940	1,386,861	1,714,860
English-Kannada	69,715	504,098	687,413
Total	2,543,340	46,030,254	36,513,563

Table 5.1 – Corpus statistics of the **complete corpus** (Collected from OPUS on August 2017) used for MT. (Tokens-En: Total number of tokens in the English side of parallel corpora. Tokens-Dr: Total number of tokens in the Dravidian language side of parallel corpora.)

	Number of sentences	Tokens-English	Tokens-Dravidian
English-Tamil	38,930	238,654	153,087
English-Telugu	38,930	238,654	164,335
English-Kannada	38,930	238,654	183,636
Total	116,790	715,962	501,058

Table 5.2 – Corpus statistics of the **multi-parallel corpus** used for MT. (Tokens-En: Total number of tokens in the English side of parallel corpora. Tokens-Dr: Total number of tokens in the Dravidian language side of parallel corpora.)

### 5.2.2 Multi-parallel Corpus

In order to enable the training of the multi-way model, we developed a **multi-parallel corpus**, which consists only of sentences that are available in all four languages. In this small subset of the complete corpus, most of the sentences for the Dravidian languages came from the translations of technical documents. The English sentences from the bilingual parallel corpora of three languages are aligned by collecting common English sentences from all three languages and their translation in the Dravidian languages. For the one-to-many multilingual models and many-to-one models (Firat et al. 2017), the parallel corpora were combined to form an English-to-Dravidian (Tamil, Telugu, and Kannada) and Dravidian-to-English NMT systems.

The corpus consists of 38,930 sentences, shown in Table 5.1. Combined, the corpus used to train multilingual NMT models consists of 116,790 sentences, 715,962 sources (English) tokens, and 501,058 target tokens.

### 5.2.3 Transliteration

In this section, we study the hypothesis of transliterating Dravidian scripts into the Latin script. Transliteration is a common method for dealing with technical terms and names while translating into another language. It is an approach where a word in one script is transformed into a different character set while attempting to maintain phonetic correspondence. As most of the Indian languages use different scripts, to take advantage of multilingual NMT models, we converted the Tamil, Telugu and Kannada script into the Latin script for a common representation before merging them into a multilingual corpus. We have used the Indic-

trans library<sup>2</sup> (Bhat et al. 2015) to transliterate the Dravidian side of the parallel corpus for three Dravidian languages, namely Tamil, Telugu, and Kannada, into the Latin script. Based on an internal evaluation campaign at FIRE2014 Shared Task on Transliterated Search, the indic-trans library produced 92.53 % accuracy for Tamil-English, 92.27 % accuracy for Telugu-English, and 91.89 % accuracy for Kannada-English.

### 5.2.4 International Phonetic Alphabet - IPA

The International Phonetic Alphabet (IPA) (International Phonetic Association 1999) contains symbols for vowels, consonants and prosodic features, such as stress and is intended to be an accurate phonetic representation for all languages. We use IPA for the phonetic transcription of Dravidian languages into a single representation. We use the EpiTran library (Mortensen et al. 2018), which is a grapheme-to-phoneme transducer supporting 61 languages. It takes the words as input and provides phonetic transcription in IPA. It has support for Tamil and Telugu but not for Kannada. Word Error Rate for the Tamil is 76.8 % and Telugu is 77.9 %. Therefore, we used the Txt2ipa<sup>3</sup> library for Kannada, which uses a dictionary mapping to convert the Kannada script into IPA script. For Txt2ipa accuracy or word error rates was not reported by authors. Figure 5.3 shows the English word *blue* in native script, transliteration and IPA. From the figure, it is clear that the transliteration has more common sub-word units than IPA.

### 5.2.5 Translation experiments

We performed our experiments with OpenNMT (Klein et al. 2017) a toolkit for neural machine translation and neural sequence modelling. After tokenization, we fed the parallel corpora to the OpenNMT pre-processing tools, i.e. OpenNMT tokenizer. Preprocessed files were then used to train the models. We used the OpenNMT parameters based on Ha et al. (2016) for training, i.e., 4 layers, 1000 for RNN size, bidirectional RNN, word level tokenization and 600-word embedding size, input feeding enabled, batch size of 64, 0.3 dropout probability and a dynamic learning rate decay<sup>4</sup>.

The approach of Ha et al. (2016) allows us to integrate the multilingual setting with a single encoder-decoder approach and without modification of the original OpenNMT model. This unified approach to extend the original NMT to multilingual NMT does not require any special treatment of the network during training. We compare the multilingual NMT model with bilingual models for both multilingual corpora and multi-way multilingual corpora. Different evaluation sets were used for testing multi-way multilingual and multilingual systems.

---

<sup>2</sup><https://github.com/libindic/indic-trans>

<sup>3</sup><https://github.com/arulalant/txt2ipa>

<sup>4</sup>This setting was used for both bilingual and multilingual NMT systems

### 5.3 Results

#### 5.3.1 Comparison of transliteration methodologies

While it is clear that IPA is generally a more fine-grained transliteration than the transliteration to Latin script, we wished to quantitatively evaluate this difference. Thus, we took the complete corpus for each language and for each character (Unicode codepoint) that occurred in the texts, we calculated its total frequency  $c_f^l$ . We then calculated the cosine similarity between the two languages,  $l_1, l_2$ , e.g.,

$$sim^{l_1, l_2} = \frac{\sum_c f_c^{l_1} f_c^{l_2}}{\sqrt{\sum_c (f_c^{l_1})^2 \sum_c (f_c^{l_2})^2}}$$

	Latin script	IPA
Tamil-Telugu	0.9790	0.7166
Tamil-Kannada	0.9822	0.5827
Telugu-Kannada	0.9846	0.8588

Table 5.3 – Cosine similarity of the transliteration of the languages under study at character level using the **complete corpus**.

	Latin script	IPA
Tamil-Telugu	0.9867	0.6769
Tamil-Kannada	0.9825	0.5602
Telugu-Kannada	0.9855	0.5679

Table 5.4 – Cosine similarity of the transliteration of the languages under study at character level using the **multi-parallel corpus**.

Table 5.3 and 5.4 shows the statistics of the cosine similarity at the character level, showing that our intuition that the Latin transliteration is much more coarse-grained is well-founded as the results show that the Latin script produces a cosine similarity of about 0.98 for these three languages whereby the IPA score is lower compared to the Latin script.

ISO 639-1	Script	Spelling	Transliteration	IPA	English
ka	Kannada	ನೀಲಿ	nili	nili	Blue
ta	Tamil	நீலம்	nilam	ni:lɑm	Blue
te	Telugu	నీలం	nilam	ni:lɑm	Blue

Figure 5.3 – Orthographic representation of word *blue* in Tamil, Telugu and Kannada shown in native script, Latin script and IPA.

To further validate this, we show in Figure 5.3 the word *blue* in all the three languages. The root word *nil* is the same in all the languages whereby Tamil and Telugu have commonality at the

whole word level. It is clear that there are far fewer commonalities in the IPA transliteration than in the Latin script transliteration.

### 5.3.2 Translation Results

Using the data, settings, and metrics described above, we investigated the impact of phonetic transcription on the machine translation of closely-related languages in multilingual NMT. We trained 54 bilingual and 18 multilingual systems corresponding to native scripts, the Latin script, and the IPA at word level and sub-word level tokenization. We use BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005) and chrF (Popović 2015) metrics for the translation evaluation. BLEU is an automatic evaluation technique which is a geometric mean of  $n$ -gram precision. It is language-independent, fast, and shows a good correlation with human judgment. It is extensively used for various MT evaluations. The METEOR metric was designed to address the drawbacks of BLEU. We also used the chrF metric to study system output at the character level, which uses F-score based on character  $n$ -grams. It is language independent and also tokenization independent.

	Native Script			Latin Script			IPA		
	B	M	C	B	M	C	B	M	C
<b>Bilingual systems results trained at word level tokenization</b>									
En-Ta	<b>40.32</b>	<b>34.79</b>	<b>62.70</b>	39.7	23.48	50.10	30.67	26.37	45.27
En-Te	20.15	21.37	40.93	<b>20.43</b>	<b>21.42</b>	<b>41.20</b>	19.3	20.06	40.09
En-Kn	<b>28.15</b>	<b>33.53</b>	<b>60.20</b>	28.13	23.46	42.96	27.11	33.50	50.78
Ta-En	<b>32.21</b>	<b>25.65</b>	<b>44.68</b>	30.72	24.78	43.60	31.2	25.29	43.60
Te-En	<b>16.24</b>	<b>28.36</b>	33.22	17.96	11.84	31.26	12.65	29.23	<b>44.01</b>
Kn-En	<b>25.93</b>	<b>22.20</b>	<b>41.88</b>	23.89	20.81	39.82	20.52	18.65	17.02
<b>Multilingual systems results trained at word level tokenization</b>									
En-Ta	43.6	34.57	64.58	<b>44.23</b>	<b>35.48</b>	<b>65.02</b>	32.94	23.86	47.03
En-Te	23.69	23.37	42.32	<b>23.98</b>	23.93	<b>42.49</b>	22.35	<b>25.98</b>	42.86
En-Kn	28.82	33.62	<b>62.73</b>	<b>31.71</b>	35.03	46.12	30.59	<b>36.45</b>	53.94
Ta-En	29.8	24.83	46.64	<b>35.66</b>	<b>28.43</b>	<b>47.44</b>	33.86	27.34	46.89
Te-En	17.82	32.34	<b>56.61</b>	<b>22.95</b>	<b>24.68</b>	36.14	16.39	24.34	48.29
Kn-En	25.11	18.50	42.60	<b>28.31</b>	<b>27.63</b>	<b>42.95</b>	24.46	24.54	19.83

Table 5.5 – BLEU (B), METEOR (M) and chrF (C) scores are illustrated for systems trained with the native script, Latin script and IPA. The native script is different for Tamil, Telugu and Kannada. Latin script and IPA are common script representations. Best results for each system are shown in bold.

### 5.3.3 Analysis of Latin script results

In order to provide a consistent evaluation of results, we compared the system outputs using the native script in all settings, instead of using the output translations in IPA and Latin script. Thus, we back-transliterated the generated translations using the Indic-trans library from

## Chapter 5. Comparison of Different Orthography

	Native Script			Latin Script			IPA		
	B	M	C	B	M	C	B	M	C
<b>Bilingual systems results trained at word level tokenization</b>									
En-Ta	31.91	22.94	43.77	<b>36.18</b>	<b>31.24</b>	<b>49.45</b>	28.67	22.92	32.35
En-Te	37.70	36.53	45.39	<b>38.67</b>	<b>34.12</b>	<b>48.44</b>	30.39	32.21	38.35
En-Kn	25.45	12.67	38.49	<b>26.51</b>	<b>28.66</b>	<b>39.87</b>	23.37	16.55	35.66
Ta-En	31.49	<b>37.61</b>	41.33	<b>34.75</b>	37.15	<b>43.24</b>	36.61	36.24	37.59
Te-En	35.30	32.23	49.35	36.44	34.69	42.72	<b>38.84</b>	<b>37.65</b>	<b>49.40</b>
Kn-En	<b>33.14</b>	21.71	44.76	30.17	<b>32.08</b>	<b>51.71</b>	24.87	18.63	45.53
<b>Multilingual system results trained at word level tokenization</b>									
En-Ta	37.32	38.94	50.56	<b>41.99</b>	<b>43.67</b>	49.11	38.45	39.66	<b>52.38</b>
En-Te	38.75	38.66	52.83	<b>39.67</b>	<b>42.75</b>	<b>56.44</b>	32.39	32.21	43.35
En-Kn	35.67	28.03	55.12	<b>37.85</b>	<b>32.43</b>	<b>60.53</b>	34.93	26.22	57.38
Ta-En	<b>36.03</b>	<b>32.32</b>	<b>54.46</b>	34.53	31.33	52.55	30.47	27.74	52.23
Te-En	34.22	31.17	53.14	<b>42.42</b>	<b>33.72</b>	<b>56.77</b>	30.72	25.82	52.28
Kn-En	32.15	46.65	59.49	<b>36.47</b>	<b>33.79</b>	<b>63.79</b>	34.59	41.06	56.12
<b>Bilingual systems results trained at sub-word level tokenization</b>									
En-Ta	36.11	20.30	53.43	<b>46.82</b>	<b>39.55</b>	<b>62.13</b>	43.63	36.36	61.90
En-Te	37.53	36.24	44.56	<b>39.47</b>	<b>36.34</b>	<b>58.45</b>	38.2	33.76	69.06
En-Kn	35.99	27.71	<b>55.37</b>	<b>39.20</b>	<b>42.94</b>	52.07	30.77	27.29	53.11
Ta-En	32.56	<b>23.42</b>	29.00	<b>36.62</b>	23.12	<b>44.35</b>	29.75	22.47	23.61
Te-En	36.12	18.93	56.63	<b>38.82</b>	<b>35.01</b>	<b>54.39</b>	39.5	25.95	37.65
Kn-En	34.85	29.26	43.86	<b>34.98</b>	<b>38.92</b>	<b>51.65</b>	33.87	24.27	45.00
<b>Multilingual systems results trained at sub-word level tokenization</b>									
En-Ta	39.25	31.91	<b>62.18</b>	<b>40.77</b>	<b>36.66</b>	56.52	31.34	27.32	52.16
En-Te	37.63	38.16	64.20	<b>38.33</b>	<b>43.34</b>	<b>67.45</b>	35.20	23.76	59.06
En-Kn	37.17	30.31	56.39	<b>37.85</b>	<b>37.08</b>	<b>59.03</b>	53.21	29.93	54.46
Ta-En	<b>37.18</b>	<b>34.69</b>	<b>57.58</b>	35.52	31.27	55.01	36.86	32.78	56.68
Te-En	<b>35.79</b>	<b>23.67</b>	46.76	29.61	23.28	46.97	28.43	20.39	<b>37.24</b>
Kn-En	<b>34.15</b>	39.84	62.19	30.53	<b>40.74</b>	<b>64.29</b>	27.36	24.56	29.38

Table 5.6 – BLEU (B), METEOR (M) and chrF (C) scores are shown for systems trained with the native script, Latin script and IPA for **multi-parallel corpora with different evaluation set**. The native script is different for Tamil, Telugu and Kannada. Latin script and IPA are common script representations. Best results for each system are shown in bold.

Latin script to native script and ran the evaluation metrics for both the corpora. Table 5.5 and 5.6 compare the results of various NMT generated translation in BLEU, METEOR, and chrF. We observe that the translations from the Latin script-based system provide an improvement in terms of BLEU, METEOR and chrF scores for translation from English to Tamil, Telugu, and Kannada for the bilingual systems for the multi-parallel corpus. This trend continues in the evaluation scores for the multilingual model as well. The multilingual systems outperform the baseline bilingual systems trained on the native script. The results are shown in Table 5.6. The METEOR and chrF score also show the same trend as the BLEU scores. Compared to the bilingual NMT system based on the native script, the multilingual NMT system based on

the Latin script has improvement in the BLEU score for translation from English to Dravidian languages.

In the other direction, i.e., from Tamil, Telugu, and Kannada to English, the results are different. The Tamil–English setting, based on the native script, has a higher BLEU score than the Latin Script for the multi-parallel corpus. For the Telugu–English setting, based on Latin script, there is an improvement in BLEU score and for the Kannada-English settings based on Latin script there is an improvement in BLEU score. The multilingual model of Tamil-English and Telugu-English have higher BLEU score based on the native script than the Latin script, except for the Kannada-English model where the Latin script-based models outperform the native script-based models. This might be caused by translating from many languages to a single language, in our case, English.

#### 5.3.4 Analysis of IPA results

To back-transcribe IPA translations into the native script, we trained an NMT system using the IPA corpus transcribed using Epitran and native script corpus as a parallel resource; this was to ensure that the comparison is fair between the different transliterations. For the IPA-Tamil (Script) system, we got a BLEU-1 score of 90.24 and 93.07 for chrF score. BLEU-1 94.11, and chrF 94.37 for IPA-Telugu. For IPA-Kannada the BLEU-1 score was 90.51, and chrF was 89.34. We then transcribed the evaluation data to a native script using the above NMT systems. Despite the promising results in multilingual NMT, IPA results are lower compared to Latin script-based systems. We observed that the scores of BLEU, METEOR, and chrF are lower than the results based on the native script in bilingual NMT translations in Table 5.5 and 5.6. It is noticeable that the scores from Dravidian languages to English trained with IPA representations did not improve the translation quality. This is due to the fact that the IPA representation is more detailed at the phonetic level than the Latin script transliteration.

#### 5.3.5 Comparing BPE with word-level models

There are two broad approaches to tokenize corpora for MT. The first approach involves word-level tokenization, and the second is sub-word level tokenization (Byte Pair Encoding). At sub-word level, closely related languages have a high degree of similarity, which makes it possible to effectively translate shared sub-words (Kunchukuttan and Bhattacharyya 2017). Byte Pair Encoding (BPE) avoids OOV issues by representing a more frequent sub-word as atomic unit (Sennrich et al. 2016b). We train our models on space-separated tokens (words) and sub-word units. Sub-word tokenization is proven to improve the results in the translation of rare and unseen words for language pairs like English–German, English–French and other languages (Sennrich et al. 2016b). Our experiments on the generated translations of the models based on the BPE corpus reveal that the systems based on Latin script have higher BLEU score in all targeted translation direction, i.e. from English to Dravidian language and vice versa. Moreover, by analyzing the METEOR and chrF scores, we note that systems based

on the Latin script using sub-word segmented corpora effectively reduce the translation errors. Again, we observed improvements from English into Dravidian languages but a drop in results for the other direction. Results for the model trained at the sub-word level are shown in Table 5.6. The transliteration-based multilingual system outperforms both the native and the IPA script-based multilingual system. These results indicated that the coarse-grained transliteration to Latin script gives an improvement of MT results by better taking advantage of closely-related languages.

**5.3.6 Error Analysis**

We observed improved performance using Latin script compared to native script and IPA, which is due to the limited number of characters, which better represents the phonological similarity of these languages. We see that the Latin transliteration mostly outperforms both the native script and the IPA transliteration and furthermore that the sub-word tokenization also improves performance. Surprisingly, the combination of these methodologies does not seem to be effective.

We can explain this by the example of the words ‘nilam’ and ‘nili’, which when we apply sub-word tokenization become ‘nil’ and ‘am’ or ‘i’. While Tamil and Telugu have similar morphology for this word, the common tokens of ‘am’ and ‘i’ are difficult to map to Kannada.

For word-level representation in the native script, the number of translation units can increase with corpus size, especially for morphologically rich languages, like Dravidian languages which lead to many OOVs. Thus, a single script with sub-word units addresses the data sparsity issue most effectively.

	<b>Ideal</b>	<b>Acceptable</b>	<b>Possibly Acceptable</b>	<b>Unacceptable</b>
<b>Native Script</b>				
En-Ta	8	11	14	17
Ta-En	8	13	18	11
<b>Transliteration</b>				
En-Ta	8	14	12	16
Ta-En	9	13	21	7
<b>IPA</b>				
En-Ta	6	14	17	13
Ta-En	3	18	18	11

Table 5.7 – Manual evaluation results of 50 sentences from testset generated by different systems for Tamil.

We performed a manual analysis of the outputs generated by the different systems. Table 5.7 show the results of manual evaluation. We used four categories based on the work by (Coughlin 2003):

**Ideal** Grammatically correct with all information accurately transferred.

**Acceptable** Comprehensible with the accurate transfer of all important information.

**Possibly Acceptable** Some information transferred accurately.

**Unacceptable** Not comprehensible and/or not much information transferred accurately.

From the manual analysis, we found that the native script and transliteration methods are more similar in terms of ideal and acceptable translation, while IPA has fewer ideal results due to errors at the character level. The unacceptable case is high in the results from native script translation due to many out of vocabulary terms. All three methods have similar numbers of acceptable and possibly acceptable cases.

## 5.4 Conclusion

In this chapter, we described our experiments on translation across different orthographies for under-resourced languages such as Tamil, Telugu, and Kannada. We show that in the Tamil, Telugu, and Kannada to English translation direction the translation quality of bilingual NMT and multilingual NMT systems improves. In order to remove the orthographic differences between languages in the same family, we performed transcription from a native script into Latin script and IPA. We demonstrated that the phonetic transcription of parallel corpora of closely-related languages shows better results than the systems trained on the native script and that the multilingual NMT with phonetic transcription to Latin script performs better than IPA transliteration. This can be explained due to the coarse-grained natures of the transliteration, which produce more similarity at the character level in the target languages, which was shown by evaluating the cosine similarity of the character frequencies. Based on results from this we create multilingual NMT systems with corpora phonetically transcribed into Latin script in the next chapter to translated wordnet glosses.



# 6 Word Sense Translation using Multilingual NMT

## 6.1 Introduction

In this chapter, we extend the result from multilingual NMT with phonetically transcribed parallel corpora to translate word sense. The method described in this chapter can be seen as an expansion of the work on elimination of code-mixed words in the parallel corpus for SMT described in Chapter 4 with addition of a phonetic transcription to multilingual NMT.

NMT has achieved rapid development in recent years; however, conventional NMT (Bahdanau et al. 2015) creates a separate MT system for each pair of languages. Creating an individual MT system for many languages is resource consuming, considering there are around 7,000 languages in the world. Ha et al. (2016) proposed an approach to extend the Bahdanau et al. (2015) architecture to a multilingual translation by sharing the entire model. The approach of shared vocabulary across multiple languages resulted in a shared embedding space. Although the results were promising, the result of the experiments was reported in highly resourced languages such as English, German, and French, but many under-resourced languages have different syntax and semantic structure to these languages. In Chapter 5, we have shown that using languages belonging to the same family with phonetic transcription to a single script improves multilingual NMT results.

Due to the lack of parallel corpora, MT systems for less-resourced languages are not readily available. We attempt to utilize Multilingual NMT, where multiple sources and target languages are trained simultaneously without changes to the network architecture. This has been shown to improve the translation quality; however, most of the under-resourced languages use different scripts which limits the application of these multilingual NMT. In order to overcome this, we transliterate the languages on the target side and bring it into a single script to take advantage of multilingual NMT for closely-related languages. Closely-related languages refer to languages that share similar lexical and structural properties due to sharing a common ancestor. Frequently, languages in contact with other language or closely-related languages like the Dravidian, Indo-Aryan, and Slavic share words from a common root (*cognates*), which are highly semantically and phonologically similar. In the scope of the wordnet creation for under-

resourced languages, combining parallel corpora from closely related languages, phonetic transcription and creating multilingual NMT systems will be shown to improve results in this chapter. The evaluation results obtained from multilingual NMT with a transliterated corpus are better than the results of SMT from Chapter 4.

In Chapter 4, which uses existing translations of wordnets in different languages to identify contextual information for wordnet senses from a large set of generic parallel corpora to evaluate and improve the wordnets for the targeted under-resourced Dravidian languages. We use this contextual information to improve the translation quality of WordNet senses. We then showed that our approach could help overcome the drawbacks of simple translations of words without context. We removed the code-mixing based on the script of the parallel corpus to reduce noise in translation. We used SMT to create bilingual MT systems for three Dravidian languages. In this chapter, we use multilingual NMT system and we transliterate the closely related language corpus into a single script to take advantage of multilingual NMT systems.

In this chapter, we look at [RQ2:]

**RQ2:** Does phonetic transcription of parallel corpora help to take advantage of resources from closely related language to improve translation quality?

This work is published as:

- Bharathi Raja Chakravarthi, Mihael Arcan and John P. McCrae “[WordNet Gloss Translation for Under-resourced Languages using Multilingual Neural Machine Translation.](#)” In proceedings of Second Workshop on Multilingualism at the intersection of Knowledge Bases and Machine Translation – co-located with MT-Summit, Dublin, Ireland (2019).

## 6.2 Our Approach

Our approach extends that of Chapter 4 and Chapter 5 by utilizing multilingual NMT with a transliterated parallel corpus of closely related languages to translate word sense for Dravidian languages. In particular, we downloaded the data, removed code-mixing and phonetically transcribed each corpus to Latin script. Two types of experiments were performed: In the first one, we removed code-mixing and compiled a multilingual corpus by concatenating the parallel corpora from three languages. In the second one, we removed code-mixing, phonetically transcribed the corpora and then compiled the multilingual corpora by concatenating the parallel corpora from three languages.

To improve the wordnets for the targeted Dravidian languages, we collected the parallel corpora from publicly available resources, we then removed the code-mixing to create a noise-free corpora. Then we transliterate the corpora and combine the corpora with a source-target token prepended to each sentence in the corpora to create a multilingual corpus. We use this

multilingual corpus to create a multilingual NMT system for Dravidian languages from English and vice-versa. We then follow the approach of [Arcan et al. \(2016b\)](#), which uses the existing translations of wordnets in other languages to identify contextual information for wordnet senses from a large set of generic parallel corpora. We use this contextual information to improve the translation quality of WordNet senses. Then we transliterated back to the original script before evaluation with previous approaches. In the following subsections, we explain our approach in detail.

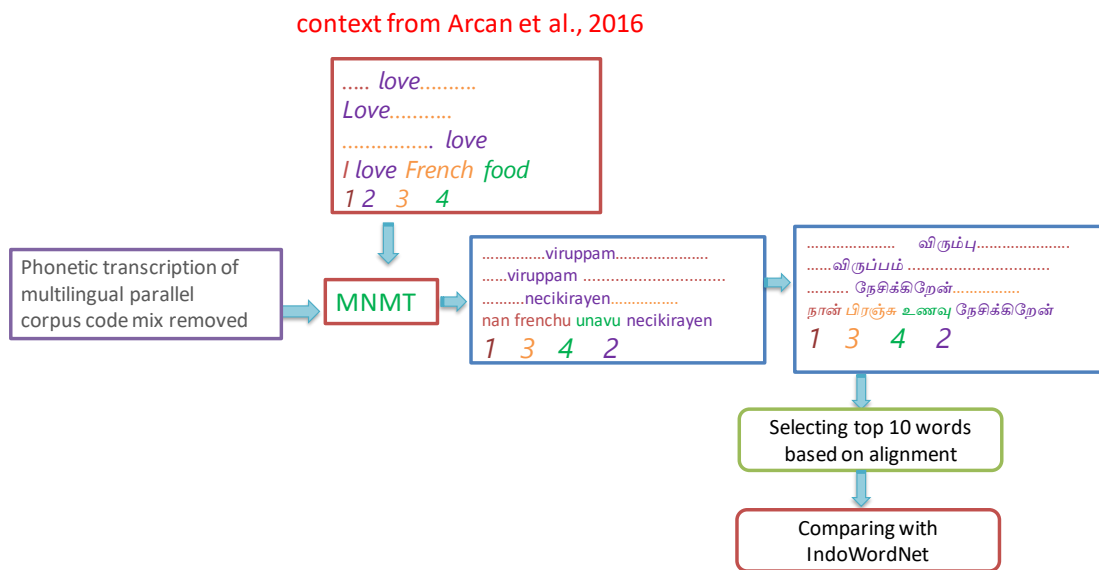


Figure 6.1 – Our approach for wordnet gloss translation using multilingual NMT

## 6.3 Experimental Setting

### 6.3.1 Multilingual NMT

[Ha et al. \(2016\)](#) and [Johnson et al. \(2017\)](#) extended the architecture of [Bahdanau et al. \(2015\)](#) to use a universal model to handle multiple source and target languages with a special tag in the encoder to determine which target language to translate. The idea is to use the unified vocabulary and training corpus without modification in the architecture to take advantage of the shared embedding. The goal of this approach is to improve the translation quality for individual language pairs, for which parallel corpus data is scarce by letting the NMT system learn the common semantics across languages and reduce the number of translation systems needed. The sentences of different languages are distinguished through language codes.

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
Number of tokens	7,738,432	6,196,245	258,165	226,264	68,197	71,697
Number of unique words	134,486	459,620	18,455	28,140	7,740	15,683
Average word length	4.2	7.0	3.7	4.8	4.5	6.0
Average sentence length	5.2	7.9	4.6	5.6	5.3	6.8
Number of sentences	449,337		44,588		13,543	

Table 6.1 – Statistics of the parallel corpora used to train the multilingual translation systems.

### 6.3.2 Data

We used the same datasets by [Chakravarthi et al. \(2018\)](#) in our experiment. We collected three Dravidian – English corpora pairs from OPUS<sup>1</sup> ([Tiedemann and Nygaard 2004](#)). Corpus statistics are shown in Table 6.1. More descriptions about the three datasets can be found in Chapter 4. We transliterated this corpus to the Latin script using the Indic-trans library<sup>2</sup>. All the sentences are first tokenized with the OpenNMT ([Klein et al. 2017](#)) tokenizer and then segmented into subword symbols using Byte Pair Encoding (BPE) ([Sennrich et al., 2016](#)). We learn the BPE merge operations across all the languages. Following [Ha et al. \(2016\)](#), we indicate the language by prepending two tokens to indicate the desired source and target language. An example of a sentence in English to be translated into Tamil would be:

```
src__en tgt__ta I like ice-cream
```

### 6.3.3 Transliteration

As the Indian languages under study are written in different scripts, they must be converted to some common representation before training the multilingual NMT to take advantage of closely-related language resources. Phonetic transcription is an approach where a word in one script is transformed into a different script by maintaining phonetic correspondence. Phonetic transcribing to Latin script and International Phonetic Alphabet (IPA) was studied in Chapter 5 and showed that Latin script outperforms IPA for the multilingual NMT Dravidian languages. The improvements in results were shown in terms of the BLEU, METEOR and chrF metric. We used Indic-trans library by [Bhat et al. \(2015\)](#), which bring all the languages into a single representation by phoneme matching algorithm. The same library can also back-transliterate from Latin script to Indian languages.

### 6.3.4 Code-Mixing

Since most of our corpora came from publicly available parallel corpora and are created by voluntary annotators or aligned automatically. Technical document translations such as

<sup>1</sup><http://opus.nlpl.eu/>

<sup>2</sup><https://github.com/libindic/indic-trans>

KDE, GNOME, and Ubuntu translations have code-mixing data since some of the technical terms may not be known to voluntary annotators for translation. But the code-mixing from OpenSubtitles are due to bilingual and historical reasons of Indian speakers (Chanda et al.; Parshad et al. 2016; 2016). Since the Internet era, English has become the international language of the younger generation. Hence, English words are frequently embedded in Indians’ speech. For our work, only intra-sentential code-mixing was taken into account. In this case, Dravidian languages as the primary languages, and English as the secondary language. We removed the English words considering only the English as a foreign word based on the script. Statistics of the removal of code-mixing are shown in Table 6.2.

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
tok	0.5% (45,847)	1.1% (72,833)	2.8% (7,303)	4.9% (12,818)	3.5% (2,425)	9.0% (6,463)
sent	0.9% (4,100)		3.1% (1,388)		3.4% (468)	

Table 6.2 – Number of sentences (sent) and number of tokens (tok) removed from the original corpus.

### 6.3.5 WordNet creation

Using contextual information to improve the translation quality of wordnet senses was shown to improve the results (Arcan et al. 2016b). The approach is to select the most relevant sentences from a parallel corpus based on the overlap of existing wordnet translations. For each synset of a wordnet entry, multiple sentences were collected that share semantic information as explained in the Section 4.2.4. We use this contextual data in English to be translated into Tamil, Telugu, and Kannada using our multilingual NMT system with attention. We used the attention vectors to get alignment of the words.

## 6.4 Results

### 6.4.1 Automatic Evaluation

We present consolidated results in Table 6.3. In addition to Precision at 1, the Table 6.3 shows Precision at 2, Precision at 5, Precision at 10. The goal of this work is to aid the human annotator in speeding up the process of wordnet creation for under-resourced languages. Precision at different levels is calculated by comparing it with IndoWordNet for the exact match out of the top 10 words from word alignment based on the attention model in multilingual NMT and alignment from SMT. The precision of all the multilingual NMT systems is greater than the baseline.

The perfect match of a word and the IndoWordNet entry is considered for precision at 1. Tamil, Telugu, and Kannada yield better precision at a different level for translation based on

## Chapter 6. Word Sense Translation using Multilingual NMT

English→Tamil				
	P@10	P@5	P@2	P@1
B-SMT	0.1200	0.1087	0.0833	0.0651
NC-SMT	0.1252	0.1147	0.0911	0.0725
NC-MNMT	<b>0.2030</b>	<b>0.1559</b>	0.1228	0.1161
NCT-MNMT	0.1816	0.1538	<b>0.1351</b>	<b>0.1320</b>
English→Telugu				
	P@10	P@5	P@2	P@1
B-SMT	0.0471	0.0455	0.0380	0.0278
NC-SMT	0.0467	0.0451	0.0382	0.0274
NC-MNMT	<b>0.0933</b>	0.0789	0.0509	0.0400
NCT-MNMT	0.0918	<b>0.0807</b>	<b>0.0599</b>	<b>0.0565</b>
English→Kannada				
	P@10	P@5	P@2	P@1
B-SMT	0.0093	0.0096	0.0080	0.0055
NC-SMT	0.0110	0.0107	0.0091	0.0067
NC-MNMT	0.0652	0.0472	0.0319	0.0226
NCT-MNMT	<b>0.0906</b>	<b>0.0760</b>	<b>0.0535</b>	<b>0.0433</b>

Table 6.3 – Results of Automatic evaluation of translated wordnet with IndoWordNet Precision at different level denoted by P@10 which means Precision at 10. B-Baseline original corpus, NC- Non-code mixed, MNMT-Multilingual Neural Machine Translation, NCT-MNMT Non Code-mixed Transliteration Multilingual Neural Machine Translation

multilingual NMT systems. For Tamil and Telugu, the translation based on multilingual NMT trained on the native script and multilingual NMT trained on the transcribed script did not have much variance. The slight reduction in the result is caused by the transliteration into and back to the original script. In the case of Kannada, which has a much lower number of parallel sentences to train compared to the other two languages, the multilingual NMT translation trained on transcribed script shows large improvement.

We have several observations. First, the precision presented is below 15 percent, and this is because these languages have minimal parallel corpora. We used the corpora collected during August 2017 from OPUS, which contains mostly translation of religious text, technical document, and subtitles. Analyzing the results by comparing with IndoWordNet is likely to be problematic since it is far from complete and is overly skewed to the classical words for these languages. Second, our method outperforms the baseline from Chapter 4 for all the languages, demonstrating the effectiveness of our framework for multilingual NMT. More importantly, transliterating the parallel corpora is more beneficial for the low resource language pair English-Kannada.

### 6.4.2 Manual Evaluation

In order to re-confirm the validity of the output in practical scenarios, we also performed a human-based evaluation in comparison with IndoWordNet entries. For human evaluation, 50 wordnet entries from wordnet were randomly selected. All these entries were evaluated according to the manual evaluation method described in [Chakravarthi et al. \(2018\)](#). The classification from Chapter 4 are explained in Subsection 4.3.2.

	B-SMT	NC-SMT	NC-MNMT	NC-MNMT-T
Agrees with IndoWordNet	18%	20%	<b>28%</b>	26%
Inflected form	12%	22%	26%	<b>30%</b>
Transliteration	4%	4%	2%	2%
Spelling variant	2%	2%	2%	2%
Correct, but not in IndoWordNet	18%	<b>24%</b>	22%	<b>24%</b>
Incorrect	46%	28%	20%	<b>16%</b>

Table 6.4 – Manual evaluation of wordnet creation for Tamil language compared with IndoWordNet (IWN) at precision at 10 presented in percentage. B-Baseline original corpus, NC-Non-code mixed, MNMT-Multilingual Neural Machine Translation, NCT-MNMT Multilingual Neural Machine Translation

Table 6.4 contains the percentage of accuracy for outputs of the wordnet translation. As mentioned earlier in Section 6.3, SMT systems trained on removing code-mixing and without removing code-mixing are used as baselines for this assessment. The baseline system shows that the cleaned data (removing code-mix) produce better results. Again, as we previously mentioned, both our multilingual NMT systems trained on cleaned data is better than the baseline system in the manual evaluation as well. From Table 6.4, we can see that there is a significant improvement over the inflected form multilingual NMT systems trained with the transcribed corpus. Perfect match with IndoWordNet is lower for multilingual NMT trained with transcribed corpus compared to multilingual NMT trained on the original script but still better than the baselines. This might be due to a back-transliteration effect. It is clear from the results that this translation can be used as an aid by annotators to create wordnets for under-resourced languages.

## 6.5 Conclusion

In this chapter, we presented how to take advantage of phonetic transcription and multilingual NMT to improve the word sense translation of under-resourced languages. The proposed approach incorporates code-mixing phenomenon into consideration as well as the phonetic transcription of closely related languages to better utilize multilingual NMT. We evaluated the proposed approach on three Dravidian languages and showed that the proposed approach outperforms the baseline by effectively leveraging the information from closely related languages. Moreover, our approach can provide better translations for a very low resourced language pair (English-Kannada). In the future, we would like to conduct an experiment by

## **Chapter 6. Word Sense Translation using Multilingual NMT**

---

transcribing the languages to one of the Dravidian languages scripts which will be able to represent information more easily than Latin script.

# 7 Multilingual Multimodal NMT utilising Phonetic Transcription

## 7.1 Introduction

In this chapter, we introduce a multilingual multimodal NMT system that utilizes orthographic information such as cognates from phonetic transcription as well as visual features. The model described in this chapter can be seen as an expansion of multilingual NMT explained in Chapter 5 to include multimodal information. Multimodal MT takes an image with a source language description then translates into a target language. The training data consists of sentence aligned parallel corpora along with their corresponding images shown in Example 7.1. Capturing rich information from multimodal content available on the Web, especially images with descriptions in English was explored in the context of NMT (Specia et al. 2016) in the WMT shared task. The development of a Multilingual Multimodal Neural Machine Translation (MMNMT) system requires multilingual parallel corpora and images which are aligned with the parallel sentences for training.

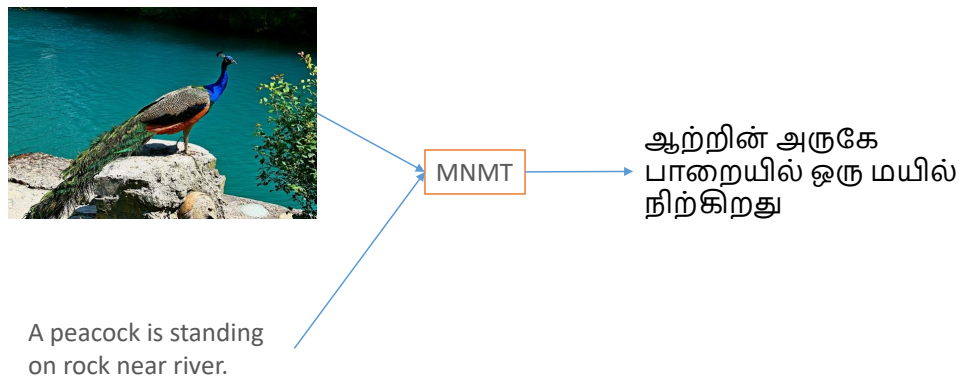


Figure 7.1 – An example of multimodal MT with an image, a description in English sentence, and translation in Tamil.

The WMT shared task provided resources for European languages namely German, Czech

## Chapter 7. Multilingual Multimodal NMT utilising Phonetic Transcription

---

and French (Elliott et al. 2017). This is the largest existing dataset at the time of writing this thesis. It contains captions, images, and translations for English, German, French and Czech and is called the Multi30K dataset which is derived from the Flickr30k dataset (Plummer et al.; Plummer et al. 2015; 2017). This task on Multimodal NMT (MNMT) was to generate image descriptions for a target language, given an image and/or a description in the source language. Typically this data was manually created with the help of bilingual annotators. Most of those datasets were expensive to construct; for example, English-German corpus was created by Elliott et al. (2016), and cost €23,000 for data collection (€0.06 per word). Such resources are not available for under-resourced languages.

In those cases, MT can be a useful tool for the quick expansion to new languages by producing candidate translations. Recent work by Dutta Chowdhury et al. (2018), carried out experiments by utilising synthetic data for Hindi-English multimodal NMT. In order to reduce the amount of time to create dataset, we pose translation as a post-editing task. We automatically translated the English sentences from the WMT corpus using a pre-trained general domain SMT system without code-mixing from Chapter 4 and NMT without code-mixing from Chapter 5.

In previous works on MNMT, the researchers utilised visual context by involving both NMT and Image Description Generation (IDG) features that explicitly uses an encoder-decoder (Cho et al. 2014). However, the encoder-decoder architecture encodes the source sentence into a fixed-length vector. To overcome this drawback (Bahdanau et al. 2015) introduced an attention mechanism to focus on parts of the source sentence. The work by Calixto and Liu (2017), carried out different experiments to incorporate visual features into NMT by projecting an image feature vector as words into the source sentence, using the image to initialise the encoder hidden state, and using image features to initialise the decoder hidden state. In Calixto et al. (2017), the authors incorporated features through a separate encoder and doubly-attentive attention of the decoder to depend on the image feature. This allowed them to predict the next word and showed that the image feature improved the translation quality. Although all these approaches have demonstrated the possibility of MNMT, they rely on manually collected corpora but under-resourced languages do not have such resources. Our work follows the doubly-attentive model (Calixto et al. 2017) with the MMDravi dataset (Chakravarthi et al. 2019c) for the multilingual model by phonetic transcription.

In Chapter 5, we have demonstrated that multilingual NMT improves translation quality for under-resourced languages. For this, we created a multilingual NMT systems without changing the architecture by introducing special tokens at the beginning of the source sentence indicating the source language and target language as shown in Figure 7.2. We follow this by introducing special tokens in the source sentence to indicate the target language.

Multilingual NMT models have been shown to increase the translation quality for under-resourced languages. Closely related Dravidian languages such as Tamil (ISO-639-1: ta), Kannada (ISO-639-1: kn), and Malayalam (ISO-639-1: ml) exhibit a large overlap in their vocabulary and strong syntactic and lexical similarities. However, the scripts used to write

Source	__opt_src__en__opt_tgt__kn a group of people standing in front of an igloo .
Target (ISO-639-1: kn)	ಇಗ್ಲೋ ಮುಂದೆ ನಿಂತಿರುವ ಜನರ ಗುಂಪು.
Source	__opt_src__en__opt_tgt__ta a group of people standing in front of an igloo .
Target (ISO-639-1: ta)	ஒரு குடிவ முன் நின்று மக்கள் குழு.
Source	__opt_src__en__opt_tgt__ml a group of people standing in front of an igloo .
Target (ISO-639-1: ml)	ഇഗ്ലൂ മുന്നിൽ നിൽക്കുന്ന ഒരു കൂട്ടം ആളുകൾ.

Figure 7.2 – Example of sentences with special tokens to indicate the source and target languages.

these languages are different and they differ in their morphology. We have shown that phonetic transcription of a corpus into Latin script improves multilingual NMT performance for under-resourced Dravidian languages in Chapter 5 and Chapter 6.

We propose to combine multilingual, phonetic transcription and multimodal content to improve the translation quality of under-resourced Dravidian languages. Our contribution in this chapter is to exploit the similar syntax and semantic structures by phonetic transcription of the corpora into Latin script along with image feature to improve the translation quality.

In this chapter, we look at the [RQ2] from a multilingual multimodal NMT point of view.

**RQ2:** Does phonetic transcription of parallel corpora help to take advantage of resources from closely related language to improve translation quality?

This work has been published at:

- Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, S Sridevy, Mihael Arcan, Manel Zarrouk, and John P. McCrae “**Multilingual Multimodal Machine Translation for Dravidian Languages utilising Phonetic Transcription.**” In proceedings of Second Workshop on Technologies for MT of Low Resource Languages – co-located with MT-Summit, Dublin, Ireland (2019).

In this chapter, we propose applying Multilingual Multimodal NMT for translating between closely related Dravidian languages and English. We created a multimodal dataset using SMT and NMT methods, trained on a general domain corpus for under-resourced languages. Combining multilingual and multimodal data along with a phonetic transcription of the corpus improves translation performance for closely related Dravidian languages.

## 7.2 Improving Multimodal NMT with Multilingual Approach

### 7.2.1 Multimodal Neural Machine Translation

The Multimodal NMT (MNMT) model proposed by [Calixto et al. \(2017\)](#) is an extension of the encoder-decoder framework which incorporates visual information. To incorporate the visual features extracted from the pre-trained model, the authors have integrated another attention mechanism to the decoder. The doubly-attentive decoder recurrent neural network is conditioned on the previous hidden state, previously emitted word, source sentence and the image via attention mechanism. In the original attention-based NMT model described in [Section 2.5](#), a single encoder for the source sentence, a single decoder for the target sentence and the attention mechanism are conditioned on the source sentence. MNMT integrates two separate attention mechanism over the source language and visual features associated with the source and target sentence. The decoder generates a target word by computing a new probability  $P(y_t = k | y_{<t}, C, A)$  given a hidden state  $s_t$ , the previously emitted word  $y_{<t}$ , and the two context vectors  $c_t$  from encoder of source sentence and  $i_t$  from image features.

$$P(y_t = k | y_{<t}, C, A) \propto \exp(L_0 \tanh(L_s s_t + L_w E_y[y_{t-1}] + L_{cs} c_t + L_{ci} i_t)) \quad (7.1)$$

$L_0, L_s, L_w, E_y, L_{cs}$ , and  $L_{ci}$  are projection matrices. The mechanism in MNMT is similar to NMT with an attention model ([Bahdanau et al. 2015](#)), except for the source sentence and previous hidden state, it also takes the context vector  $i_t$  from the image features using a double attention layer to calculate the current hidden state. The doubly-attentive model according to [Calixto et al. \(2017\)](#), calculates the time-dependent vector  $i_t$  as follows:

$$i_t = \beta_t \sum_{l=1}^L \alpha_{t,l}^{img} a_l \quad (7.2)$$

Where,

$$\beta_t = \sigma(W_\beta s_{t-1} + b_\beta) \quad (7.3)$$

$\beta_t$  is a gating scalar between  $[0, 1]$  and used to weigh the expected influence of the image context vector in relation to the next target word at time step  $t$ . The expected alignment vector of image is given by

$$\alpha_{t,l}^{img} = \frac{\exp(e_{t,l}^{img})}{\sum_{j=1}^L \exp(e_{t,j}^{img})} \quad (7.4)$$

$$e_{t,l}^{img} = (V_a^{img})^T \tanh(U_a^{img} s_t + W_a^{img} a_l) \quad (7.5)$$

$V_a^{img}$ ,  $U_a^{img}$  and  $W_a^{img}$  are model parameters. This visual attention computes a time-dependent image context vector  $i_t$  given a hidden state proposal  $s'_t$  and the image annotation vectors  $A = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L)$  using the "soft" attention.

### 7.2.2 Multilingual Multimodal Neural Machine Translation

Since we translate between closely related languages and English, we set up the translation setting in two scenarios, 1) One-to-Many and 2) Many-to-One.

#### One-to-Many Approach

In this setting, we create a model to translate from English into Tamil, Malayalam, and Kannada. The source language sentence was replicated three times for the three languages with a token indicating the target language. Figure 7.2 shows the example of sentences.

#### Many-to-One Approach

In the many-to-one MMNMT system, we create a model to translate from Tamil, Malayalam, and Kannada (Dravidian languages) to English. We replicated the English sentence three times for three languages on the target side of the corpus. We then train the MNMT system with a visual feature for individual language level with the MMDravi data. We compared the results with the MMNMT for one-to-many and many-to-one models.

## 7.3 Experimental Settings

### 7.3.1 Data

The images required for our work were collected from Flickr by [Plummer et al. \(2015\)](#). The Multi30K dataset contains parallel corpora for English and German. There were two types of multilingual annotations in the Multi30K dataset ([Elliott et al. 2016](#)). The first one is an English description for each image and its German translation. The second is a corpus of five independently collected English and German description pairs for each image. Synthetic data or back-transliterated data ([Xia et al. 2019](#)) have been widely used to improve the performance of NMT and MNMT. To produce a target side description of an image, we create a general domain SMT and NMT systems for English-Tamil, English-Kannada, and English-Malayalam pairs. We collected general domain parallel corpora for the Dravidian languages from OPUS ([Tiedemann and Nygaard 2004](#)) and [Chakravarthi et al. \(2018\)](#). The corpus statistics are shown in Table 7.1. The corpus is tokenised and standardised to lowercase. The general domain SMT system was created with Moses ([Koehn et al. 2007](#)) while the NMT system was trained with OpenNMT ([Klein et al. 2017](#)). After tokenisation, we fed the parallel corpora to Moses and OpenNMT. Preprocessed files are then used to train the models. We used the default

## Chapter 7. Multilingual Multimodal NMT utilising Phonetic Transcription

---

OpenNMT parameters for training, i.e. 2 layers LSTM with 500 hidden units for both, the encoder and decoder.

<b>Lang pair</b>	<b>Corpus Statistics</b>		
	<i>sent</i>	<i>s-tokens</i>	<i>t-tokens</i>
En-Ta	0.8M	6.4M	13.3M
En-Kn	0.5M	2.6M	4.5M
En-Ml	1.4M	16.7M	23.5M

Table 7.1 – Statistics of the parallel corpora used to train the general domain translation systems. *sent*: Number of sentences, *s-tokens*: Number of source tokens, and *t-tokens*: Number of target tokens.

The SMT and NMT systems results on the general domain evaluation dataset are shown in Table 7.2. The development and test set of the multimodal corpus was collected with the help of volunteer annotators. To reduce the annotation time, we posed the translation task of the development and test set as a post-editing task. We provided the candidate translation of the English sentence from SMT, NMT, and an option to choose the best translation or provide an original translation. Eighteen annotators participated in this annotation process, with different backgrounds, they all are native speakers of the language that they annotated. The data for the Malayalam language was collected from three different native speakers. Ten Tamil native speakers participated in creating data for the Tamil language and five Kannada native speakers annotated for the Kannada language. Since voluntary annotators are scarce and annotate little data, each sentence was annotated by only one annotator. We then selected the system that performed better based on the choice of annotators. We designed an annotation tool to meet the objective of method. We decided to use Google Forms to collect the data from the voluntary annotators. An example is shown in Figure 7.3. We chose NMT and used the general domain NMT to post-edit the translation for the training set of MMDravi.

<b>Lang pair</b>	<b>BLEU Score</b>	
	<i>SMT</i>	<i>NMT</i>
En-Ta	30.29	35.52
En-Kn	28.81	26.86
En-Ml	36.73	38.56

Table 7.2 – Results of general domain SMT and NMT translation systems on general domain evaluation set

For our tasks, all descriptions in English were converted to lowercase and tokenised, while we skipped the lowercasing step for the Dravidian languages, as these languages do not case characters. We tokenised the Dravidian language using the OpenNMT tokeniser with *segment alphabet* options for Tamil, Kannada, and Malayalam. For the sub-word level representation, we chose the 10,000 most frequent units to train the BPE (Sennrich et al. 2016b) model. We used this model for the sub-word level segmentation for the training, development, and

Choose the best translation



\*

a man sleeping in a green room on a couch .

- ஒரு படுக்கையில் ஒரு பச்சை அறையில் ஒரு மனிதன் தூங்கி.
- ஒரு மனிதன் படுக்கையில் ஒரு படுக்கையில் ஒரு green அறைக்குள் .
- NONE

Your answer

Figure 7.3 – Example of sentence and image with candidate translation to choose.

evaluation set. We trained the MMNMT model to translate from English into Dravidian languages as well as from Dravidian languages into English. Visual features were extracted from publicly available pre-trained CNNs. Specifically, we extract spatial image features using the VGG-19 network (Simonyan and Zisserman 2014). In our experiment, we pass all the images in our dataset through the pre-trained VGG-19 layered network to extract global information and use them in a separate visual attention mechanism as described in Calixto et al. (2017).

## 7.4 Results

Table 7.3 provides the BLEU scores for the MMNMT model. We observed that the translation performance of MMNMT is higher compared to the Bilingual Multimodal NMT model in BLEU. Translation from Dravidian to English has the highest improvement in terms of BLEU Score. Our experiments show that the MMNMT system compared with the bilingual system has an improvement in several language directions, which are likely gained from phonetic transcription, image features, and transfer of parameters from different languages.

Lang pair	BLEUScore		
	Baseline	MMNMT	MMNMT-T
En-Ta	50.2	51.0	<b>52.3</b>
En-Ml	35.6	36.0	<b>36.5</b>
En-Kn	44.5	45.1	<b>45.9</b>
Ta-En	45.2	47.4	<b>48.9</b>
Ml-En	34.3	36.2	<b>37.6</b>
Kn-En	50.0	50.2	<b>50.8</b>

Table 7.3 – Results are expressed in BLEU score: Baseline is Multimodal NMT, MMNMT is trained on native script, and MMNMT-T is trained utilizing phonetic transcription.

The results show that MMNMT-T with phonetically transcribed corpora helps more when translating from Dravidian languages into English than vice versa. An explanation for this is that in the dataset, each source sentence has three targets, which encourages the language model to improve the translation results. In Table 7.3, we compare the BLEU scores with a baseline approach and our method. In order to evaluate the effectiveness of our proposed model, we have explored MMNMT trained on original scripts and MMNMT trained on a single script. Our empirical results show that the best result is achieved when we phonetically transcribed the corpus and brought it to a single script for both English to Dravidian and Dravidian to English translation tasks.


	src	a black dog runs on green grass with a toy in his mouth .
	ref	ஒரு கருப்பு நாய் வாயில் ஒரு பொம்மையுடன் பச்சை புல்லில் ஓடுகிறது.
	MMNMT	ஒரு கருப்பு நாய் வாயில் ஒரு பொம்மையுடன் பச்சை புல் மீது இயங்கும் .
	MMNMT-T	ஒரு கருப்பு நாய் வாயில் ஒரு பொம்மையுடன் பச்சை புல் மீது ஓடுகிறது.

Figure 7.4 – Example showing improvement of translation quality and readability of the translation over baseline model. Errors are shown in red color.

Figure 7.4 shows the examples of where the MMNMT model improves the translation quality and readability of the translation over the baseline model. The results given by the human evaluation confirm the results observed by using evaluation BLEU metric. The second example for English-Tamil translation of the MMNMT system outperforming the baseline is shown in Figure 7.5. The example in Figure 7.4 shows an almost perfect translation obtained with the MMNMT system for English to Tamil. In the second example, the translation obtained with the MMNMT system is acceptable with the accurate transfer of important information (Coughlin 2003). This suggests the synthetic data produced by MMNMT model can be used in


	src	a woman and two men , that are dressed professionally, are having a discussion.
	ref	ஒரு பெண் மற்றும் இரண்டு ஆண்கள், தொழில்முறை உடையணிந்து, ஒரு விவாதத்தில் உள்ளார்கள்.
	MMNMT	ஒரு பெண் , மற்றும் இரு ஆண்கள் professionally ஆயத்தம்பண்ணி , ஒரு விவாதத்திலை வேண்டுகோளம் .
	MMNMT -T	ஒரு பெண் மற்றும் இரண்டு ஆண்கள், professional உடையணிந்து, ஒரு விவாதத்தில் உள்ளார்.

Figure 7.5 – Example showing translation with accurate transfer of important information. Errors are shown in red.

an under-resourced language setting to improve the translation quality.

## 7.5 Conclusion

We introduced a new dataset, named MMDravi and proposed a MMNMT method for closely related Dravidian languages to overcome the resource issues. Compared to the baseline approach, the results show that our approach can improve translation quality, especially for Dravidian languages. Our evaluation, using phonetic transcription, multilingual and multimodal NMT, has shown that the proposed MMNMT-T outperforms the existing approach of multimodal, multilingual in low-resource NMT across all the language pairs considered.

## 7.6 Summary

This study makes several contributions to the field. First, we did a comparative analysis of different orthographies for under-resourced MT. Secondly, we created MT systems for under-resourced languages by transcribing the parallel corpora into a common script (Latin script). This shows improvement over the SMT results from Chapter 4 for word sense translation. Finally, we also created a multilingual MNMT utilising phonetic transcription. We observe that the phonetic transcription to a single script improves translation quality in multilingual NMT for all three cases.

In this part, we described an efficient approach for RQ2 in multilingual NMT. Our focus was on the most recent NMT with attention model. We gave a detailed analysis for proposed methods, and compared these approach with state-of-the-art results. In particular, we demonstrated the efficiency of phonetic transcription on both small and large training data using multilingual NMT.

However, some of the languages do not have enough resources to create multilingual NMT systems such as Tulu (ISO 639-3: tcy). For such languages, monolingual resources can be

## **Chapter 7. Multilingual Multimodal NMT utilising Phonetic Transcription**

---

found on the Internet. Therefore, the next step is to create a word-level translation using bilingual lexicon induction methods. In the next part of this thesis, we switch our attention to bilingual lexicon induction with orthographic information.

**Orthography in Bilingual Lexicon** **Part V**  
**Induction**



# 8 Bilingual Lexicon Induction across Orthographically-distinct Languages

## 8.1 Introduction

In the previous two parts, we presented our work for **RQ1** and **RQ2** in MT scenarios. The assumption we made in these scenarios is that we have some amount of parallel corpora available for the languages under study. However, this is not always true for all under-resourced languages. On the other hand, we have a monolingual corpora on the Internet for these under-resourced languages. Such monolingual corpora can be utilised to create bilingual lexicons.

Bilingual lexicon induction (BLI) is the process of creating lexicons for two or more languages that share the collective meaning of words 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 since it reduces time. However, sentence-aligned parallel data is not available for all languages, and it is costly to create. Methods based on unsupervised or semi-supervised learning can utilise readily available monolingual data to induce bilingual lexicons. Artetxe et al. (2018a) showed that an iterative self-learning method could bootstrap 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. Nevertheless, BLI is a challenging task for under-resourced languages due to lack of seed dictionaries and large monolingual corpora. We propose to use the IndoWordNet as a seed dictionary for the languages under study.

BLI between closely-related languages has been shown to perform better than between unrelated languages. Since, closely related languages often share many linguistics properties, including similar words sharing a meaning called cognates as we discussed in Part IV. Cognates are words which have a similar meaning and similarity in orthography based on etymological relationships (Kondrak et al. 2003). Computational models of monolingual embeddings also exhibit similarities across closely related languages (Mikolov et al. 2013b) based on the as-

## Chapter 8. Bilingual Lexicon Induction across Orthographically-distinct Languages

---

sumption that word embeddings in different languages have approximately the same structure. This isomorphic property was exploited by [Artetxe et al. \(2018a\)](#) and [Lample et al. \(2018\)](#) to map monolingual word embeddings in different languages to 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. As it has been shown that even small seed information can improve the performance over fully unsupervised methods ([Patra et al. 2019](#)), the usage of such alignments is likely to improve performance. Previous works used 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. In this part of the thesis, we look at **RQ3**:

**RQ3:** Does orthographic information help to improve embedding-based bilingual lexicon induction (BLI) for closely-related languages?

The goal of this chapter is to utilise orthographic information between languages which 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 the closely related languages. This chapter has two principal contributions: Firstly, we show that the use of the IndoWordNet as a seed dictionary validates our cognate property of closely related languages. Secondly, recall that in [Part IV](#), we have shown that transliteration of parallel corpora for multilingual NMT improves the translation quality. Based on this, we look into the use of transliteration, and we show that it is an effective and necessary step which yields more isomorphic embeddings and obtains more robust BLI.

### 8.2 Our Approach

For MT systems, it is essential to have a corpus of written documents, as well as well-defined lexicons and grammar for the selected languages. Hence in this chapter, we will focus on the four chosen Dravidian languages namely Tamil, Telugu, Kannada, and Malayalam which are spoken by approximately 210 million people ([Steever 2015](#)) across the world either as a first language or second language.

Even though these languages share many words with 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 language without being translated ([Burrow and Emeneau 1961](#)). [Figure 8.1](#) also shows the words for ‘salt’, ‘coin’, ‘vehicle’, ‘walk’ are similar in both Tamil and Malayalam. The languages Tamil, Malayalam and Telugu have their own written script symbols whereas Telugu and Kannada have significant similarities in their script symbols. Though Telugu and Kannada

have similarities in their script symbols, they are not readily intelligible for speakers of another language, and as both scripts are represented with a different Unicode set, they cannot be properly recognised by a machine.

ISO 639-1	ta	ml	ka	te	en
Script	Tamil	Malayalam	Kannada	Telugu	Latin
	உப்பு (Uppu)	ഉപ്പു (Uppu)	ಉಪ್ಪು (Uppu)	ఉప్పు (Upp)	Salt
	நாணயம் (Nanayam)	നാണയം (Nanayam)	ನಾಣ್ಯ (Nanya)	నాణం (Nanem)	Coin
	வாகனம் (Vakanam)	വാഹനം (Vahanam)	ವಾಹನ (Vahana)	వాహనం (Vahanam)	Vehicle
	நடை (Nadai)	നട (Nada)	ನಡಿ (Nadi)	నడిచి (Nadici)	Walk

Figure 8.1 – Example of cognate words in Dravidian language

Since the languages evolved sharing geographical, etymological and political borders, the cognates may have evolved similar meanings or borrowed words from each other. The example ‘salt’ and ‘walk’ can be traced back to proto-Dravidian language, however ‘coin’ and ‘vehicle’ are borrowed from the Indo-European language due to geographical borders. In Chapter 5, we have compared the Latin script and International Phonetic Alphabet (IPA) for multilingual translation system and showed that bringing the Dravidian languages into the Latin script outperforms the multilingual NMT system in native script and IPA. Inspired by this we bring the Dravidian language monolingual corpora into a single script (Latin script) to take advantage of cognate property.

### 8.2.1 Orthographic Information in BLI

Riley and Gildea (2018) proposed two methods to utilise orthographic information to improve the BLI, as explained in Section 2.6. The first method is an orthographic extension of word embeddings, where each word in 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:

$$A = A_{\text{source}} \cup A_{\text{target}} \tag{8.1}$$

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

$$O_{ij} = c_e \cdot \text{count}(A_j, w_i), O \in \{O_{\text{source}}, O_{\text{target}}\} \tag{8.2}$$

Then embedding matrices were concatenated with orthographic matrices as below:

$$X' = [X; O_{\text{source}}], \quad Z' = [Z; O_{\text{target}}] \tag{8.3}$$

## Chapter 8. Bilingual Lexicon Induction across Orthographically-distinct Languages

---

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

$$X''_{i*} = \frac{X'_{i*}}{\|X'_{i*}\|}, \quad Z''_{i*} = \frac{Z'_{i*}}{\|Z'_{i*}\|} \quad (8.4)$$

These new matrices are used in place of  $X$  and  $Z$  in the self-learning process of Section 2.6.

Traditional frameworks use the seed dictionary to learn the linear mapping  $W$  between  $X$  and  $Z$  given the seed dictionary and uses it to induce dictionary  $D$ , however it suffers from the non-availability of a seed dictionary. Artetxe et al. (2017) proposed self-learning

---

### Algorithm 1 Self-learning framework

---

**Input:**  $X$  (source embeddings)

**Input:**  $Z$  (target embeddings)

**Input:**  $D$  (seed dictionary)

1: **repeat**

2:  $W \leftarrow \text{LEARN-MAPPING}(X, Z, D)$

3:  $D \leftarrow \text{LEARN-DICTIONARY}(X, Z, W)$

4: **until** convergence criterion

5:  $\text{EVALUATE-DICTIONARY}(D)$

---

As discussed in Section 2.6, the goal of  $\text{LEARN-MAPPING}(X, Z, D)$  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 minimized as explained in Equation 2.4 and Equation 2.5. The  $X''$  and  $Z''$  are new matrices from Equation 8.4 are used in the place of  $X$  and  $Z$  to include orthographic information.

$\text{LEARN-DICTIONARY}(X, Z, W)$  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  $j = \text{argmax}_k (X_{i*}W) \cdot Z_{k*}$  and  $D_{ij}$  is set to 0 otherwise as explained in Section 2.6.

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

$$\text{Translation Accuracy} = \frac{\text{Correct Translations}}{\text{All Samples}} \quad (8.5)$$

where the 'correct translation' is the number of correct target words in the test set, 'all samples' are the total number of samples in the test set.

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 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 is denoted as NL.

$$\text{NL}(w_1, w_2) = \frac{L(w_1, w_2)}{\max(|w_1|, |w_2|)} \quad (8.6)$$

Orthographic similarity of two words  $w_1$  and  $w_2$  would be  $\log(2.0 - \text{NL}(w_1, w_2))$ . The edit distance for a subset of possible word pairs is only considered as the vast majority of word pairs are orthographically very dissimilar, resulting in a normalised edit distance close to 1 and an orthographic similarity close to 0.

### 8.2.2 Longest Common Subsequence

Levenshtein distance is a common 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. In this chapter, we explore the potential use of the Longest Common Subsequence (LCS) for BLI. LCS ([Paterson and Dančik 1994](#)) is a similarity measure of two or more strings to find the longest subsequence common to all sequences in two or more strings. LCS was used to extract morphological variations and their lexicons generations from monolingual corpora ([Hulden et al.; Sorokin 2014; 2016](#)). LCS was also used to identify cognate candidates during the construction of N-best translation lexicons from parallel text ([Melamed 1995](#)), and for the automatic evaluation of translation quality ([Lin and Och 2004](#)).

A sequence  $Z = [z_1, z_2, \dots, z_n]$  is a subsequence of another sequence  $X = [x_1, x_2, \dots, x_m]$ , if there exists a sequence  $[i_1, i_2, \dots, i_k]$  of indices of  $X$  such that for all  $j = 1, 2, \dots, k$ , where  $x_{i_j} = z_j$ . Given two sequences  $X$  and  $Y$ , the LCS of  $X$  and  $Y$  is a common subsequence with maximum length. More formally

$$\text{LCS}(X_i, Y_j) = \begin{cases} \emptyset & \text{if } i = 0 \text{ or } j = 0 \\ \text{LCS}(X_{i-1}, Y_{j-1})' x_i & \text{if } i, j > 0 \text{ and } x_i = y_j \\ \max\{\text{LCS}(X_i, Y_{j-1}), \text{LCS}(X_{i-1}, Y_j)\} & \text{if } i, j > 0 \text{ and } x_i \neq y_j \end{cases} \quad (8.7)$$

In previous works, [Artetxe et al. \(2018a\)](#) uses the dot product of two word 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 dictionary induction phase of self-learning framework. LCS

## Chapter 8. Bilingual Lexicon Induction across Orthographically-distinct Languages

is used to measure the orthographic similarity of languages (Melamed; Nakov and Ng 1995; 2012). Since Dravidian languages have rich morphological features, this will be beneficial in extracting their cognate information by LCS. To include LCS, we modify Riley and Gildea (2018) similarity score for each word pair during the dictionary induction phase. The normalised edit distance of Equation 8.6 is modified as below:

$$\text{NLCS}(w_1, w_2) = 1 - \frac{\text{LCS}(w_1, w_2)}{\max(|w_1|, |w_2|)} \quad (8.8)$$

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

Language Pair (ISO 639-1)	Word Pair	NLCS	NL
kn-ml	hajaradant-hajarulla	0.4545	0.6363
kn-ml	rahasyadaan-rahasyadan	0.2307	0.3076
kn-ta	navratn-navmani	0.3750	0.5000
kn-ta	tandeilladant-tacoppanillat,	0.4285	0.7857
kn-te	poojaniyavadantah-poojyaniyulu	0.5555	0.6666
kn-te	atyagatyavadant-atyavasaramin	0.5000	0.5625
ml-ta	navaratnam-navmani	0.5454	0.6363
ml-ta	tatanillat-tacoppanillat	0.3571	0.4285
ml-te	navaratnam-navratnal	0.2727	0.3636
ml-te	tatanillat-tandriless	0.5454	0.7272
ta-te	tacoppanillat-tandriless	0.6428	0.7857
ta-te	sammadam-samardhinchada	0.6000	0.6666

Table 8.1 – Examples from training set for comparison of NLCS and NL

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 for words. 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 in Chapter 6 (Chakravarthi et al. 2019b) shown improvement in the translation of word sense and compare the results with IndoWordNet. In Chapter 5, we (Chakravarthi et al. 2019a) showed that Latin script outperforms the International Phonetic Alphabet for Multilingual NMT for Dravidian languages. Inspired by this, we used the Indic-trans library by Bhat et al. (2015) to transliterate the corpus. We show examples comparing of NLCS and NL between cognate words in Table 8.2.

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.; Berg-Kirkpatrick et al.; Hauer et al. 2011; 2010; 2017). Hauer et al. (2017) created a seed dictionary based on the cognates of the related languages using orthographic information. They have shown that approaches which include orthographic information outperform the

	Walk			Salt		
Language Pair		NLCS	NL		NLCS	NL
ta-te	nadai-nadici	0.33	0.33	uppu-uppu	0.25	0.25
ta-kn	nadai-nadi	0.19	0.20	uppu-uppu	0.00	0.00
ta-ml	nadai-nada	0.19	0.20	uppu-uppu	0.00	0.00
te-kn	nadici-nadi	0.33	0.33	upp-uppu	0.25	0.25
te-ml	nadici-nada	0.50	0.50	upp-uppu	0.25	0.25
kn-ml	nadi-nada	0.25	0.25	uppu-uppu	0.00	0.00
	Coin			Vehicle		
Language Pair		NLCS	NL		NLCS	NL
ta-te	nanayam-nanem	0.42	0.42	vakanam-vahanam	0.14	0.14
ta-kn	nanayam-nanya	0.28	0.28	vakanam-vahana	0.28	0.28
ta-ml	nanayam-nanayam	0.00	0.00	vakanam-vahanam	0.14	0.14
te-kn	nanem-nanya	0.40	0.40	vahanam-vahana	0.14	0.14
te-ml	nanem-nanayam	0.42	0.42	vahanam-vahanam	0.00	0.00
kn-ml	nanya-nanayam	0.28	0.28	vahana-vahanam	0.14	0.14

Table 8.2 – Comparison of NLCS and NL between languages for example of cognate words

previous approaches for closely related languages. Previous works did not study languages which have different scripts and used Levenstein distance without considering morphological properties.

### 8.2.3 Data

Lexicons such as WordNet for English or EuroWordNet for European languages have been used to improve MT quality (Kazemi et al. 2016). EuroWordNet is 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, no such resources are available, 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 from IndoWordNet to one-to-one word mapping by replicating the source word multiple times. Table 8.3 shows the seed dictionary statistics and Table 8.4 show the statistics for the test set. Test set was randomly chosen among the mapping from IndoWordNet for languages under study. Even though IndoWordNet is not perfect as we discussed in Chapter 4 and Chapter 6, this is the one resource readily available for the under-resourced Dravidian language experiments.

Words that share a similar context are semantically related. Based on this, word embedding

Language Pair	Number of entries
ta-te	11,666
ta-kn	4,353
ta-ml	18,731
te-kn	12,769
te-ml	3,791
kn-ml	4,639

Table 8.3 – Number of entries in the initial bilingual lexicons used as a seed dictionary for the experiment

Language Pair	Number of entries
ta-te	1,982
ta-kn	1,930
ta-ml	1,918
te-kn	2,000
te-ml	2,000
kn-ml	1,999

Table 8.4 – Number of entries in the test set for the experiment

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 (Mikolov et al.; Mikolov et al. 2013a; 2013b), global word representation-based models (GloVe) (Pennington et al. 2014) and FastText (Bojanowski et al.; Grave et al. 2017; 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.

Wikipedia is a free online encyclopedia which is created by volunteers from a different region of the world. It has contents in more than 200 languages. Wikipedia dumps (Wikidump)<sup>1</sup> for Tamil, Telugu, Malayalam and Kannada were downloaded to create a monolingual embedding for each language. Wikidumps were downloaded from July 2019. Wikiextractor<sup>2</sup> 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 8.5. 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. These 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

<sup>1</sup><https://dumps.wikimedia.org/>

<sup>2</sup><https://github.com/attardi/wikiextractor>

Gildea (2018). We used the indic-trans library<sup>3</sup> 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 8.5 – Number of sentences and number of tokens extracted from wikimedia.org for Dravidian languages

### 8.2.4 Experiment Settings

Given that the main focus of our **RQ3** is on bringing closely related languages into a single script, we transcribed the 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. Similar to Riley and Gildea (2018), we stop training when the improvement on the average cosine distance for the induced dictionary is below  $1 \times 10^{-6}$  between succeeding 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.

## 8.3 Results and Discussion

We show results of all eight cases studied for BLI in Table 8.6. First, we added the seed dictionary created from IndoWordNet in (Artetxe et al. 2017) while maintaining the corpora in native script. Adding the seed dictionary showed small improvement over the baseline. Then we did an experiment with the methodology proposed by Riley and Gildea (2018) as another baseline still maintaining the corpora in the native script, which showed improvement over the baseline. In our approach, we first transliterate the corpora with numerals only and a seed dictionary. In the final experiment, we used LCS to improve the baseline methodology with transliteration numerals only and a seed dictionary.

Once the dictionary (D) is learned after the self-learning process. We use D to create translations for source words from the test set, and then it is compared with the gold set (target words) 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 8.5. The total number of cases examined for language pairs under study is given in Table 8.4.

<sup>3</sup><https://github.com/libindic/indic-trans>

## Chapter 8. Bilingual Lexicon Induction across Orthographically-distinct Languages

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 Tamil-Malayalam language pair yield 220 word translations correctly and total number of samples were 2,000. Hence, translation accuracy is 11%. The results of experiments indicate accuracy improvements with a seed dictionary and transliteration. We further investigate our result to explicate the effects of cognates from similar languages.

		ta-te	ta-kn	ta-ml	te-kn	te-ml	kn-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
Our approach	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
Our approach LCS	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 8.6 – Performance comparison of bilingual lexicon induction on test data for Dravidian languages. Translation accuracy is represented in percentage. N: native script, T: transliteration, seed-dic: seed dictionary.

As can be seen from Table 8.6, our approach with LCS outperforms baseline methods within their groups for all four languages in six pairs. Moreover, transliteration gives the best accuracy across all baseline methods for all six language pairs, and this can be explained from the examples in Figure 8.1. Our method with a transliteration and LCS outperforms all bases on Tamil-Telugu. Interestingly, the transliteration and LCS fails on Malayalam-Kannada with numerals seed dictionary. Since the monolingual corpora for these languages are very small compared to the high resource languages such as English, Spanish, and German studied by Artetxe et al. (2018c). For example, 1,384,170,636+65,648,657,780 tokens for German, 702,638,442+36,237,951,419 tokens for Italian, and 127,176,620+6,059,887,126 tokens for Finnish from Wikipedia and Common Crawl respectively used in training word vector, which are comparatively very high to number for tokens in Dravidian languages as shown in Table 8.5. 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 resources 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 this setting that our method really stands out.

As it can be seen, our method with LCS obtains best results in all language pairs and directions, with highest 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 embedding mapping between languages which uses different scripts.

All approaches do better with the transliterated 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 unsupervised with only numbers as seed dictionary. When the word vectors are not rich enough, the baseline methods fail entirely to map the embeddings without 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 the seed dictionary.

## 8.4 Conclusion

In this chapter, 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. We have also proposed to use LCS during the dictionary induction phase and have shown better performance. Our findings strengthen the idea that cognate information and bringing closely related languages into a single script improves the performance of constructing bilingual lexicons. In the next chapter, we conclude the thesis and give future directions.



## **Conclusions and Future Work** **Part VI**



## 9 Conclusions and Future work

In this chapter, we provide conclusions for the previous chapters and revisit the research questions with the answers we have provided to them. We then summarise the contributions of our work in this thesis. Later in this chapter, we explore the various possibilities for future research.

### 9.1 Conclusion and Research Questions Answered

Traditional MT algorithms require a large amount of sentence aligned parallel corpora to learn representations at the sub-word, word, and sentence level. However, under-resourced languages lack such resources. This thesis aimed to investigate these challenges and to make advances towards utilising orthographic information to improve MT for under-resourced languages. We considered three research questions to perform this

**RQ1:** Does removing foreign words (code-mix) based on orthography improve word sense translation quality?

**RQ2:** Does phonetic transcription of parallel corpora help to take advantage of resources from closely related language to improve translation quality?

**RQ3:** Does orthographic information help to improve embedding-based bilingual lexicon induction for closely-related languages?

We started the thesis by illustrating the benefits of removing code-mixing based on the orthography in the parallel corpora. We explained the attempt to create and expand wordnets for Dravidian languages such as Tamil, Kannada and Telugu. We used the parallel corpora available from multiple sources, such as OPUS, to build MT frameworks to translate the word senses in the Princeton WordNet into the under-resourced languages mentioned above. The OPUS parallel corpora had several code-mixing points. By excluding them depending on the orthography, the treatment of code-mixing improves the quality of the results outperforming

## Chapter 9. Conclusions and Future work

---

the baseline results of wordnet entries. This was demonstrated in Chapter 4 of Part III in answer to **RQ1**. Therefore we conclude that the removal of foreign words based on orthography improves translation quality for a specific task such as word sense translation. This work still applies with limited effort to resource development of under-resourced languages.

However, MT is an active research area in which technologies are being developed expeditiously. With the availability of powerful computing, neural network-based MT became the state-of-the-art training algorithm in MT. In Part IV, we turned our attention to phonetic transcription in multilingual NMT, to study our second research question **RQ2**.

In Chapter 5, we described our experiments on translation across different orthographies for under-resourced languages such as Tamil, Telugu, and Kannada. We show that in this Tamil, Telugu, and Kannada to English translation direction, the translation quality of bilingual NMT and multilingual NMT systems improves. In order to remove the orthographic differences between languages in the same family, we performed transcription from a native script into the Latin script and IPA. We demonstrated that the phonetic transcription of parallel corpora of closely-related languages shows better results than the system trained on the native script and that multilingual NMT with phonetic transcription to Latin script performs better than the IPA transliteration. This can be explained due to the coarse-grained nature of the transliteration, which produces more similarity at the character level in the target languages, which we proved by evaluating the cosine similarity of the character frequencies. Based on results from this we create multilingual NMT systems with corpora phonetically transcribed into Latin script in the chapters to translate word sense.

Revisiting the state-of-the-art and the experimental results from Chapter 4, we observed in Chapter 5 that the multilingual NMT performs better than bilingual NMT when the parallel corpora are phonetically transcribed to a single script. In Chapter 6, we presented how to take advantage of phonetic transcription and multilingual NMT to improve word sense translation of under-resourced languages. The proposed approach brings code-mixing into consideration as well as the phonetic transcription of closely related language to better utilise multilingual NMT. We evaluated the method on three Dravidian languages and showed that the proposed approach outperforms the baseline by efficiently leveraging the information from closely related languages. Moreover, our approach can provide better translations for a very low resourced language pair (English-Kannada).

We introduced a new dataset (MMDravi) and proposed a MMNMT approach for closely related Dravidian languages to overcome the issue of the languages being under-resourced. Compared to the baseline approach, the results show that our approach can improve translation quality, especially for Dravidian languages. Our evaluation, using phonetic transcription, multilingual and multimodal NMT, has demonstrated that the proposed MMNMT-T outperforms the existing approach of multimodal, multilingual low-resource NMT across all the language pairs considered in Chapter 7.

Our results from Chapter 5, Chapter 6, and Chapter 7 indicate that the systems trained on pho-

netically transcribed corpora performed better than on the native script, which is supported by the observation and cosine similarity scores of the multilingual parallel corpora.

In Part V, we addressed **RQ3**. In Chapter 8, we have explored bringing closely related languages into a single script feed the task of bilingual lexicon induction from monolingual word embeddings. We created a seed dictionary and test set for Tamil, Telugu, Malayalam and Kannada from the IndoWordNet. We have also proposed to use LCS during the dictionary induction phase and have shown better performance. Our findings strengthen the idea that cognate information and bringing closely related languages into a single script improves the performance on the bilingual lexicon induction task.

## 9.2 Contributions

In this thesis, we have investigated novel approaches for utilising orthographic information to improve MT of under-resourced languages. The contributions of our work can be summarised as follows.

- **Code-Mixing.** We presented an approach to remove code-mixing based on orthography to improve the translation quality of under-resourced languages. By doing so, we have created clean parallel corpora for three Dravidian languages, namely, Tamil-English, Telugu-English, and Kannada-English. We used these clean corpora to create SMT systems and translated the wordnet gloss, which was used to improve IndoWordNet.
- **Phonetic Transcription.** This study makes several contributions to this thesis. First, we carried out a comparative analysis of different orthographies for under-resourced MT. Secondly, we created MT systems for under-resourced by transcribing the parallel corpora into a common script (Latin script) based on the comparative analysis. This shows improvement over the SMT results from Chapter 4 for wordnet gloss translation. Finally, we also create multilingual MNMT utilising phonetic transcription. We observe that the phonetic transcription to a single script improves translation quality in multilingual NMT for all three cases. In this part, we described an efficient approach for RQ2 in the multilingual NMT. Our focus was on the most recent neural network training from NMT. We gave a detailed analysis of proposed methods, and compared these approaches with state-of-the-art results. In particular, we demonstrated the efficiency of phonetic transcription on both small and large training data using multilingual NMT. We also introduced the MMDravi dataset for multimodal multilingual NMT.
- **Orthographic Information in BLI.** We have explored bringing closely related languages into a single script and their impact on the task of bilingual lexicon induction from monolingual word embeddings. We designed a BLI dictionary induction phase to use LCS, which is linguistically sound compared with the linguistically unsound Levenshtein distance approach. We also proposed to use existing resources such as IndoWordNet entries as a seed dictionary and test set for the under-resourced Dravidian languages.

Our findings strengthen the idea that cognate information and bringing closely related languages into a single script improve the performance of mining bilingual lexicons.

### 9.3 Future Work

We believe our work achieved all the three research questions stated. The achievements open up new research questions and point to interesting future work:

- There are several possible extensions to the approach presented in this thesis. We compared the orthography between native script, Latin script, and IPA. In future work, we would like to conduct an experiment by transcribing the languages to one of the Dravidian languages scripts which will be able to represent information more easily than Latin script.
- NMT suffers from out-of-vocabulary (OOV) words for under-resourced languages. We plan to explore different methods to fuse bilingual lexicon into an end-to-end state-of-the-art NMT system to overcome OOV.
- Multimodal machine translation is the task of translating from a source text into the target language using information from other modalities than text such as an image. Other modalities associated with text usually add new information, which helps to ground the meaning of the corresponding text. We plan to release multilingual translations as an addition to the Flickr30k dataset and explore the effect of the quality of this synthetic data in our future work.
- Under-resourced languages lack language resources in terms of a text corpus. However, there are plenty of videos available for many under-resourced languages. Nevertheless, it remains our future work to explore a way to collect the e-text corpus along with videos from social media and create multimodal resources for under-resourced languages.

## Bibliography

- Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. **Multi-dialect neural machine translation and dialectometry**. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics, Hong Kong. <https://www.aclweb.org/anthology/Y18-1001>.
- Steven Abney and Steven Bird. 2010. **The Human Language Project: Building a Universal Corpus of the World's Languages**. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 88–97. <http://www.aclweb.org/anthology/P10-1010>.
- Roe Aharoni, Melvin Johnson, and Orhan Firat. 2019. **Massively multilingual neural machine translation**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 3874–3884. <https://doi.org/10.18653/v1/N19-1388>.
- Benyamin Ahmadnia, Javier Serrano, and Gholamreza Haffari. 2017. **Persian-Spanish low-resource statistical machine translation through English as pivot language**. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., Varna, Bulgaria, pages 24–30. <https://www.aclweb.org/anthology/R17-1004/>.
- Iñaki Alegria, Xabier Artola, Arantza Diaz De Ilarraza, and Kepa Sarasola. 2011. **Strategies to develop language technologies for less-resourced languages based on the case of Basque**. In *Proceedings of 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. pages 42–46. <https://artxiker.ccsd.cnrs.fr/artxibo-00783393/document>.
- Mihael Arcan, Mauro Dragoni, and Paul Buitelaar. 2016a. **Translating ontologies in real-world settings**. In *Proceedings of the 15th International Semantic Web Conference (ISWC-2016)*. Kobe, Japan. <https://www.insight-centre.org/content/translating-ontologies-real-world-settings>.
- Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2016b. **Expanding wordnets to new languages with multilingual sense disambiguation**. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 97–108.

## Bibliography

---

- <https://www.aclweb.org/anthology/C16-1010>.
- 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. <https://www.aclweb.org/anthology/D16-1250/>.
- 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)*. Association for Computational Linguistics, Vancouver, Canada, pages 451–462. <https://doi.org/10.18653/v1/P17-1042>.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. **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)*. Association for Computational Linguistics, Melbourne, Australia, pages 789–798. <https://doi.org/10.18653/v1/P18-1073>.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. **Unsupervised statistical machine translation**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 3632–3642. <https://doi.org/10.18653/v1/D18-1399>.
- 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*. Association for Computational Linguistics, Florence, Italy, pages 5002–5007. <https://doi.org/10.18653/v1/P19-1494>.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.
- Moses Omoniyi Ayeomoni. 2006. Code-switching and code-mixing: Style of language use in childhood in Yoruba speech community. *Nordic Journal of African Studies* 15(1):90–99.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. **Neural machine translation by jointly learning to align and translate**. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. <http://arxiv.org/abs/1409.0473>.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, pages 65–72. <http://www.aclweb.org/anthology/W05-0909>.
- Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for Arabic dialects that utilizes multitask learning (MTL). *Computational intelligence and neuroscience* 2018.

- Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014a. **Code mixing: A challenge for language identification in the language of social media**. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics, Doha, Qatar, pages 13–23. <https://doi.org/10.3115/v1/W14-3902>.
- Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014b. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*. pages 13–23.
- Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. pages 883–891.
- Yonatan Belinkov and Yonatan Bisk. 2018. **Synthetic and natural noise both break neural machine translation**. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=BJ8vJebC->.
- Alevtina Bemova, Karel Oliva, and Jarmila Panevova. 1988. **Some Problems of Machine Translation Between Closely Related Languages**. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*. <http://www.aclweb.org/anthology/C88-1010>.
- 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*. Association for Computational Linguistics, Los Angeles, California, pages 582–590. <https://www.aclweb.org/anthology/N10-1083>.
- N. Bertoldi, R. Zens, M. Federico, and W. Shen. 2008. **Efficient speech translation through confusion network decoding**. *IEEE Transactions on Audio, Speech, and Language Processing* 16(8):1696–1705. <https://doi.org/10.1109/TASL.2008.2002054>.
- Kamaddev Bhanuprasad and Mats Svenson. 2008. **Errgrams - A Way to Improving ASR for Highly Inflected Dravidian Languages**. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*. <http://www.aclweb.org/anthology/I08-2113>.
- Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. **IIT-H system submission for FIRE2014 shared task on transliterated search**. In *Proceedings of the Forum for Information Retrieval Evaluation*. ACM, New York, NY, USA, FIRE '14, pages 48–53. <https://doi.org/10.1145/2824864.2824872>.
- Riyaz Ahmad Bhat, Irshad Ahmad Bhat, Naman Jain, and Dipti Misra Sharma. 2016. **A house united: Bridging the script and lexical barrier between Hindi and Urdu**. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. pages 397–408. <http://aclweb.org/anthology/C/C16/C16-1039.pdf>.
- Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources

## Bibliography

---

- Association (ELRA), Valletta, Malta.
- Pushpak Bhattacharyya, Mitesh M. Khapra, and Anoop Kunchukuttan. 2016. **Statistical machine translation between related languages**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, San Diego, California, pages 17–20. <https://doi.org/10.18653/v1/N16-4006>.
- Alexandra Birch, Barry Haddow, Ivan Tito, Antonio Valerio Miceli Barone, Rachel Bawden, Felipe Sánchez-Martínez, Mikel L. Forcada, Miquel Esplà-Gomis, Víctor Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Wilker Aziz, Andrew Secker, and Peggy van der Kreeft. 2019. **Global under-resourced media translation (GoURMET)**. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. European Association for Machine Translation, Dublin, Ireland, pages 122–122. <https://www.aclweb.org/anthology/W19-6723>.
- Michael Bloodgood and Benjamin Strauss. 2017. **Acquisition of translation lexicons for historically unwritten languages via bridging loanwords**. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, pages 21–25. <https://doi.org/10.18653/v1/W17-2504>.
- 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.
- Eric Brill and Robert C. Moore. 2000. **An improved error model for noisy channel spelling correction**. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Hong Kong, pages 286–293. <https://doi.org/10.3115/1075218.1075255>.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. **The mathematics of statistical machine translation: Parameter estimation**. *Computational Linguistics* 19(2):263–311. <https://www.aclweb.org/anthology/J93-2003>.
- 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). Trubner.
- Iacer Calixto and Qun Liu. 2017. **Incorporating global visual features into attention-based neural machine translation**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 992–1003. <http://aclweb.org/anthology/D17-1105>.
- Iacer Calixto, Qun Liu, and Nick Campbell. 2017. **Doubly-attentive decoder for multi-modal neural machine translation**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1913–1924. <https://doi.org/10.18653/v1/P17-1175>.

- Michael Carl. 2000. **A model of competence for corpus-based machine translation**. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*. <https://www.aclweb.org/anthology/C00-2145>.
- 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*. The Global WordNet Conference 2018 Committee.
- 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)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, volume 70 of *OpenAccess Series in Informatics (OASICs)*, pages 6:1–6:14. <https://doi.org/10.4230/OASICs.LDK.2019.6>.
- 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*. European Association for Machine Translation, Dublin, Ireland, pages 1–7. <https://www.aclweb.org/anthology/W19-7101>.
- 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*. European Association for Machine Translation, Dublin, Ireland, pages 56–63. <https://www.aclweb.org/anthology/W19-6809>.
- Joyce Y. C. Chan, Houwei Cao, P. C. Ching, and Tan Lee. 2009. **Automatic Recognition of Cantonese-English Code-Mixing Speech**. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009*. <https://www.aclweb.org/anthology/O09-5003>.
- Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. 2016. **Columbia-jadavpur submission for EMNLP 2016 code-switching workshop shared task: System description**. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics, Austin, Texas, pages 112–115. <https://doi.org/10.18653/v1/W16-5814>.
- Yu Chen and Tania Avgustinova. 2019. **Machine translation from an intercomprehension perspective**. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Association for Computational Linguistics, Florence, Italy, pages 192–196. <https://doi.org/10.18653/v1/W19-5425>.
- Colin Cherry and Hisami Suzuki. 2009. **Discriminative Substring Decoding for Transliteration**. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1066–1075. <http://www.aclweb.org/anthology/D09-1111>.
- Rohan Chitnis and John DeNero. 2015. **Variable-length word encodings for neural translation**

## Bibliography

---

- models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 2088–2093. <https://doi.org/10.18653/v1/D15-1249>.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. **Learning phrase representations using RNN encoder–decoder for statistical machine translation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pages 1724–1734. <https://doi.org/10.3115/v1/D14-1179>.
- Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. **Improving statistical machine translation accuracy using bilingual lexicon extraction with paraphrases**. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. Department of Linguistics, Chulalongkorn University, Phuket, Thailand, pages 262–271. <https://www.aclweb.org/anthology/Y14-1032>.
- Kenneth Ward Church. 1993. **Char-align: A program for aligning parallel texts at the character level**. In *31st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Columbus, Ohio, USA, pages 1–8. <https://doi.org/10.3115/981574.981575>.
- Alina Maria Ciobanu and Liviu P. Dinu. 2014. **Automatic detection of cognates using orthographic alignment**. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 99–105. <https://doi.org/10.3115/v1/P14-2017>.
- Trevor Cohn and Mirella Lapata. 2007. **Machine translation by triangulation: Making effective use of multi-parallel corpora**. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 728–735. <https://www.aclweb.org/anthology/P07-1092>.
- Marta R Costa-Jussa, Mireia Farrús, José B Marino, and José AR Fonollosa. 2012. Study and comparison of rule-based and statistical catalan-spanish machine translation systems. *Computing and informatics* 31(2):245–270.
- Marta R. Costa-jussà and José A. R. Fonollosa. 2016. **Character-based neural machine translation**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 357–361. <https://doi.org/10.18653/v1/P16-2058>.
- Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *In Proceedings of MT Summit IX*. pages 63–70.
- Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text. the last language identification frontier? volume 54, pages 41–64.
- Eva Dauphin and Veronika Lux. 1996. **Corpus-based annotated test set for machine translation evaluation by an industrial user**. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. <https://www.aclweb.org/anthology/C96-2188>.

- Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. **Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach**. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pages 131–140. <https://www.aclweb.org/anthology/W18-3817>.
- Mona Diab, Julia Hirschberg, Pascale Fung, and Tamar Solorio. 2014. **Proceedings of the First Workshop on Computational Approaches to Code Switching**. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics. <http://www.aclweb.org/anthology/W14-3900>.
- Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. **A call for prudent choice of subword merge operations in neural machine translation**. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. European Association for Machine Translation, Dublin, Ireland, pages 204–213. <https://www.aclweb.org/anthology/W19-6620>.
- Qing Dou, Ashish Vaswani, and Kevin Knight. 2014. **Beyond parallel data: Joint word alignment and decipherment improves machine translation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 557–565. <https://doi.org/10.3115/v1/D14-1061>.
- Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. **Multimodal neural machine translation for low-resource language pairs using synthetic data**. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*. Association for Computational Linguistics, Melbourne, pages 33–42. <https://doi.org/10.18653/v1/W18-3405>.
- 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*. Association for Computational Linguistics, USA, HLT '11, page 409–419.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. **Understanding back-translation at scale**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 489–500. <https://doi.org/10.18653/v1/D18-1045>.
- Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*. European Language Resources Association (ELRA), pages 2868–2872.
- Ahmed El Kholy and Nizar Habash. 2012. **Orthographic and morphological processing for English-Arabic statistical machine translation**. *Machine Translation* 26(1–2):25–45. <https://doi.org/10.1007/s10590-011-9110-0>.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. **Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description**. In *Proceedings of the Second Conference on Machine Translation*. Association

## Bibliography

---

- for Computational Linguistics, pages 215–233. <http://aclweb.org/anthology/W17-4718>.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. **Multi30K: Multilingual English-German Image Descriptions**. In *Proceedings of the 5th Workshop on Vision and Language*. Association for Computational Linguistics, pages 70–74. <https://doi.org/10.18653/v1/W16-3210>.
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. **Data augmentation for low-resource neural machine translation**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 567–573. <https://doi.org/10.18653/v1/P17-2090>.
- Marzieh Fadaee and Christof Monz. 2018. **Back-translation sampling by targeting difficult words in neural machine translation**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 436–446. <https://doi.org/10.18653/v1/D18-1040>.
- Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.
- Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, Springer, pages 231–243.
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. **Multi-way, multilingual neural machine translation with a shared attention mechanism**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 866–875. <https://doi.org/10.18653/v1/N16-1101>.
- Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. 2017. **Multi-way, multilingual neural machine translation**. *Comput. Speech Lang.* 45(C):236–252. <https://doi.org/10.1016/j.csl.2016.10.006>.
- Andrea Fischer, Klára Jágrová, Irina Stenger, Tania Avgustinova, Dietrich Klakow, and Roland Marti. 2016. Orthographic and morphological correspondences between related slavic languages as a base for modeling of mutual intelligibility. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. pages 4202–4209.
- Lluís Formiga and Jose A. R. Fonollosa. 2012. **Dealing with input noise in statistical machine translation**. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, Mumbai, India, pages 319–328. <https://www.aclweb.org/anthology/C12-2032>.
- Victoria Fromkin, Robert Rodman, and Nina Hyams. 2018. *An Introduction to Language*. Cengage Learning.
- Pascale Fung. 1995. **Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus**. In *Third Workshop on Very Large Corpora*. <https://www.aclweb.org/anthology/W95->

0114.

- Igal Golan, Shalom Lappin, and Mori Rimón. 1988. **An active bilingual lexicon for machine translation**. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*. <https://www.aclweb.org/anthology/C88-1042>.
- Charlotte Gooskens. 2007. **The contribution of linguistic factors to the intelligibility of closely related languages**. *Journal of Multilingual and Multicultural Development* 28(6):445–467. <https://doi.org/10.2167/jmmd511.0>.
- Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. **Generalizing back-translation in neural machine translation**. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy, pages 45–52. <https://doi.org/10.18653/v1/W19-5205>.
- 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)*.
- Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pages 439–448.
- Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. **Cognate-aware morphological segmentation for multilingual neural translation**. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 386–393. <https://doi.org/10.18653/v1/W18-6410>.
- Francisco Guzman, Houda Bouamor, Ramy Baly, and Nizar Habash. 2016. **Machine translation evaluation for Arabic using morphologically-enriched embeddings**. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1398–1408. <https://www.aclweb.org/anthology/C16-1132>.
- Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the International Workshop on Spoken Language Translation*.
- Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. **Learning bilingual lexicons from monolingual corpora**. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, pages 771–779. <https://www.aclweb.org/anthology/P08-1088>.
- Jan Hajic. 2000. **Machine translation of very close languages**. In *Sixth Applied Natural Language Processing Conference*. Association for Computational Linguistics, Seattle, Washington, USA, pages 7–12. <https://doi.org/10.3115/974147.974149>.
- Jack Halpern. 2018. **Very large-scale lexical resources to enhance Chinese and Japanese machine translation**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA),

## Bibliography

---

- Miyazaki, Japan. <https://www.aclweb.org/anthology/L18-1137>.
- Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2018. **Collection and analysis of code-switch Egyptian Arabic-English speech corpus**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. <https://www.aclweb.org/anthology/L18-1601>.
- 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*. Association for Computational Linguistics, Valencia, Spain, pages 619–624. <https://www.aclweb.org/anthology/E17-2098>.
- Auður Hauksdóttir. 2014. **An Innovative World Language Centre : Challenges for the Use of Language Technology**. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA). <http://www.aclweb.org/anthology/L14-1618>.
- Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 187–197.
- Csilla Horváth, Ágoston Nagy, Norbert Szilágyi, and Veronika Vincze. 2016. Where bears have the eyes of currant: Towards a Mansi WordNet. In *Proceedings of the Eighth Global WordNet Conference*. pages 130–134.
- 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*. Association for Computational Linguistics, Gothenburg, Sweden, pages 569–578. <https://doi.org/10.3115/v1/E14-1060>.
- International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.
- Ann Irvine and Chris Callison-Burch. 2017. **A comprehensive analysis of bilingual lexicon induction**. *Computational Linguistics* 43(2):273–310. <https://www.aclweb.org/anthology/J17-2001>.
- Saurav Jha, Akhilesh Sudhakar, and Anil Kumar Singh. 2019. Learning cross-lingual phonological and orthographic adaptations: a case study in improving neural machine translation between low-resource languages. *Journal of Language Modelling* 7(2):101–142.
- Robert Jimerson and Emily Prud'hommeaux. 2018. ASR for Documenting Acutely Under-Resourced Indigenous Languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association

- (ELRA), Miyazaki, Japan.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. **Google’s multilingual neural machine translation system: Enabling zero-shot translation**. *Transactions of the Association for Computational Linguistics* 5:339–351. <https://www.aclweb.org/anthology/Q17-1024>.
- Hiroyuki Kaji. 1988. **An efficient execution method for rule-based machine translation**. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*. <https://www.aclweb.org/anthology/C88-2167>.
- Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. **Neural machine translation for low-resource languages without parallel corpora**. *Machine Translation* 32(1):167–189. <https://doi.org/10.1007/s10590-017-9203-5>.
- Arefeh Kazemi, Antonio Toral, and Andy Way. 2016. Using wordnet to improve reordering in hierarchical phrase-based statistical machine translation .
- Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI’16, page 2741–2749.
- Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. **Pivot-based transfer learning for neural machine translation between non-English languages**. 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)*. Association for Computational Linguistics, Hong Kong, China, pages 866–876. <https://doi.org/10.18653/v1/D19-1080>.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. **OpenNMT: Open-Source Toolkit for Neural Machine Translation**. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, pages 67–72. <http://www.aclweb.org/anthology/P17-4012>.
- Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. **Toward statistical machine translation without parallel corpora**. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, pages 130–140. <https://www.aclweb.org/anthology/E12-1014>.
- Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT.
- Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. **Moses: Open Source**

## Bibliography

---

- 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*. Association for Computational Linguistics, pages 177–180. <http://www.aclweb.org/anthology/P07-2045>.
- Philipp Koehn and Kevin Knight. 2000a. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press, page 711–715.
- Philipp Koehn and Kevin Knight. 2000b. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press, page 711–715.
- 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. <https://www.aclweb.org/anthology/N03-2016>.
- Steven Krauwer. 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003* pages 8–15.
- Taku Kudo. 2018. **Subword regularization: Improving neural network translation models with multiple subword candidates.** In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 66–75. <https://doi.org/10.18653/v1/P18-1007>.
- Taku Kudo and John Richardson. 2018. **SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.** In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, pages 66–71. <https://doi.org/10.18653/v1/D18-2012>.
- Arun Kumar, Ryan Cotterell, Lluís Padró, and Antoni Oliver. 2017. **Morphological Analysis of the Dravidian Language Family.** In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 217–222. <http://aclweb.org/anthology/E17-2035>.
- Adimugan Kumaran and Tobias Kellner. 2007. A generic framework for machine transliteration. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 721–722.
- Anoop Kunchukuttan and Pushpak Bhattacharyya. 2017. **Learning variable length units for SMT between related languages via Byte Pair Encoding.** In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. Association for Computational Linguistics, pages 14–24. <http://aclweb.org/anthology/W17-4102>.
- Anoop Kunchukuttan, Mitesh Khapra, Gurmeet Singh, and Pushpak Bhattacharyya. 2018. **Leveraging orthographic similarity for multilingual neural transliteration.** *Transactions of the As-*

- sociation for Computational Linguistics 6:303–316. <https://www.aclweb.org/anthology/Q18-1022>.
- Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, and Pushpak Bhattacharyya. 2017. **Utilizing lexical similarity between related, low-resource languages for pivot-based SMT**. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 283–289. <https://www.aclweb.org/anthology/I17-2048>.
- Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. **Neural machine translation into language varieties**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Brussels, Belgium, pages 156–164. <https://doi.org/10.18653/v1/W18-6316>.
- 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*. <https://openreview.net/forum?id=H196sainb>.
- Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. *Machine Translation Summit XIV* page 83.
- Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. **Fully character-level neural machine translation without explicit segmentation**. *Transactions of the Association for Computational Linguistics* 5:365–378. <https://www.aclweb.org/anthology/Q17-1026>.
- Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.
- William Lewis, Robert Munro, and Stephan Vogel. 2011. **Crisis MT: Developing a cookbook for MT in crisis situations**. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 501–511. <https://www.aclweb.org/anthology/W11-2164>.
- Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. **Findings of the first shared task on machine translation robustness**. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy, pages 91–102. <https://doi.org/10.18653/v1/W19-5303>.
- Zhenhao Li and Lucia Specia. 2019. **Improving neural machine translation robustness via data augmentation: Beyond back-translation**. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, Hong Kong, China, pages 328–336. <https://doi.org/10.18653/v1/D19-5543>.
- 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*

## Bibliography

---

- Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, USA, ACL '04, page 605–es. <https://doi.org/10.3115/1218955.1219032>.
- Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. Truecasing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 152–159.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. **Effective approaches to attention-based neural machine translation**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421. <https://doi.org/10.18653/v1/D15-1166>.
- Mike Maxwell and Baden Hughes. 2006. **Frontiers in Linguistic Annotation for Lower-Density Languages**. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*. Association for Computational Linguistics, pages 29–37. <http://www.aclweb.org/anthology/W06-0605>.
- I. Dan Melamed. 1995. **Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons**. In *Third Workshop on Very Large Corpora*. <https://www.aclweb.org/anthology/W95-0115>.
- I. Dan Melamed. 1999. **Bitext maps and alignment via pattern recognition**. *Computational Linguistics* 25(1):107–130. <https://www.aclweb.org/anthology/J99-1003>.
- Mohamed Amine Menacer, David Langlois, Denis Jouvét, Dominique Fohr, Odile Mella, and Kamel Smâïli. 2019. Machine translation on a parallel code-switched corpus. In *Canadian Conference on Artificial Intelligence*. Springer, pages 426–432.
- 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*. <http://arxiv.org/abs/1301.3781>.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. **Exploiting similarities among languages for machine translation**. *CoRR* abs/1309.4168. <http://arxiv.org/abs/1309.4168>.
- George A. Miller. 1995. **WordNet: A lexical database for english**. Association for Computing Machinery, New York, NY, USA, volume 38, page 39–41. <https://doi.org/10.1145/219717.219748>.
- Zhang Min, Li Haizhou, and Su Jian. 2004. Direct orthographical mapping for machine transliteration. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 716.
- Gaurav Mohanty, Abishek Kannan, and Radhika Mamidi. 2017. Building a sentiwordnet for Odia. *EMNLP 2017* page 143.
- David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources*

- and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France.
- Andrea Mulloni and Viktor Pekar. 2006. **Automatic detection of orthographics cues for cognate recognition**. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy. <https://www.aclweb.org/anthology/L06-1420/>.
- Preslav Nakov. 2008. **Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing**. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Columbus, Ohio, pages 147–150. <https://www.aclweb.org/anthology/W08-0320>.
- Preslav Nakov and Hwee Tou Ng. 2009. **Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages**. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1358–1367. <http://www.aclweb.org/anthology/D09-1141>.
- 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.
- Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 301–305.
- Graham Neubig and Junjie Hu. 2018. **Rapid adaptation of neural machine translation to new languages**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 875–880. <https://doi.org/10.18653/v1/D18-1103>.
- Xing Niu, Michael Denkowski, and Marine Carpuat. 2018. **Bi-directional neural machine translation with synthetic parallel data**. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics, Melbourne, Australia, pages 84–91. <https://doi.org/10.18653/v1/W18-2710>.
- Mai Oudah, Amjad Almahairi, and Nizar Habash. 2019. **The impact of preprocessing on Arabic-English statistical and neural machine translation**. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. European Association for Machine Translation, Dublin, Ireland, pages 214–221. <https://www.aclweb.org/anthology/W19-6621>.
- 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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. <https://doi.org/10.3115/1073083.1073135>.
- Rana D. Parshad, Suman Bhowmick, Vineeta Chand, Nitu Kumari, and Neha Sinha. 2016. **What is India speaking? Exploring the “Hinglish” invasion**. *Physica A: Statistical Mechanics and its Applications* 449:375 – 389. <https://doi.org/https://doi.org/10.1016/j.physa.2016.01.015>.

## Bibliography

---

- Mike Paterson and Vlado Dančik. 1994. Longest common subsequences. In Igor Prívvara, Branislav Rován, and Peter Ruzička, editors, *Mathematical Foundations of Computer Science 1994*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 127–142.
- 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*. Association for Computational Linguistics, Florence, Italy, pages 184–193. <https://doi.org/10.18653/v1/P19-1018>.
- Jasabanta Patro, Bidisha Samanta, Saurabh Singh, Abhipsa Basu, Prithwish Mukherjee, Monojit Choudhury, and Animesh Mukherjee. 2017. **All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2264–2274. <https://doi.org/10.18653/v1/D17-1240>.
- 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)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. <https://doi.org/10.3115/v1/D14-1162>.
- Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multi-lingual database. In *Proc. 1st Int'l Conference on Global WordNet*.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. **Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models**. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Washington, DC, USA, ICCV '15, pages 2641–2649. <https://doi.org/10.1109/ICCV.2015.303>.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. **Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models**. *International Journal of Computer Vision* 123(1):74–93. <https://doi.org/10.1007/s11263-016-0965-7>.
- Maja Popović. 2015. **chrF: character n-gram F-score for automatic MT evaluation**. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 392–395. <https://doi.org/10.18653/v1/W15-3049>.
- Maja Popović, Mihael Arcan, and Filip Klubička. 2016. **Language Related Issues for Machine Translation between Closely Related South Slavic Languages**. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*. The COLING 2016 Organizing Committee, pages 43–52. <http://www.aclweb.org/anthology/W16-4806>.
- Maja Popović and Nikola Ljubešić. 2014. **Exploring cross-language statistical machine translation for closely related south Slavic languages**. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*. Association for Computational Linguistics, Doha, Qatar, pages 76–84. <https://doi.org/10.3115/v1/W14->

4210.

- Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 401–409.
- P. Prakash and R. Malatesha Joshi. 1995. *Orthography and Reading in Kannada: A Dravidian Language*, Springer Netherlands, Dordrecht, pages 95–108.
- S Rajendran, S Arulmozi, B Kumara Shanmugam, S Baskaran, and S Thiagarajan. 2002. Tamil wordnet. In *Proceedings of the First International Global WordNet Conference*. Mysore: CIIL. pages 271–274.
- S Rajendran, G Shivapratap, V Dhanlakshmi, and KP Soman. 2010. Building a wordnet for Dravidian languages. In *Proceedings of the Global WordNet Conference (GWC 10)*. Citeseer. Citeseer.
- Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for English-Tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*. pages 113–122.
- Prakash Ranjan, Bharathi Raja, Ruba Priyadharshini, and Rakesh Chandra Balabantaray. 2016. **A comparative study on code-mixed data of Indian social media vs formal text**. In *2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, pages 608–611. <https://ieeexplore.ieee.org/document/7918035>.
- Sujith Ravi and Kevin Knight. 2011. **Deciphering foreign language**. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 12–21. <https://www.aclweb.org/anthology/P11-1002>.
- 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*. NLP Association of India, Trivandrum, India, pages 71–78. <https://www.aclweb.org/anthology/W15-5910>.
- Shruti Rijhwani, Royal Sequiera, Monojit Choudhury Choudhury, and Kalika Bali. 2016. Translating codemixed tweets: A language detection based system. In *3rd Workshop on Indian Language Data Resource and Evaluation-WILDRE-3*. pages 81–82.
- 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)*. Association for Computational Linguistics, Melbourne, Australia, pages 390–394. <https://doi.org/10.18653/v1/P18-2062>.
- Rashed Rubby Riyadh and Grzegorz Kondrak. 2019. Joint approach to deromanization of code-mixed texts. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*. pages 26–34.
- Michael Rosner and Kurt Sultana. 2014. Automatic methods for the extension of a bilingual dic-

## Bibliography

---

- tionary using comparable corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pages 3790–3797.
- M. Schuster and K. Nakajima. 2012. **Japanese and Korean voice search**. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pages 5149–5152. <https://doi.org/10.1109/ICASSP.2012.6289079>.
- Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. **Multi-lingual unsupervised NMT using shared encoder and language-specific decoders**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 3083–3089. <https://doi.org/10.18653/v1/P19-1297>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. **Improving neural machine translation models with monolingual data**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 86–96. <https://doi.org/10.18653/v1/P16-1009>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. **Neural Machine Translation of Rare Words with Subword Units**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1715–1725. <https://doi.org/10.18653/v1/P16-1162>.
- Hendra Setiawan, Haizhou Li, Min Zhang, and Beng Chin Ooi. 2005. Phrase-based statistical machine translation: A level of detail approach. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *Natural Language Processing – IJCNLP 2005*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 576–587.
- Michel Simard, George F Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*. IBM Press, pages 1071–1082.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.
- Thamar Solorio and Yang Liu. 2008. **Learning to Predict Code-Switching Points**. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 973–981. <http://www.aclweb.org/anthology/D08-1102>.
- Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. **Code-switching for enhancing NMT with pre-specified translation**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 449–459. <https://doi.org/10.18653/v1/N19-1044>.
- 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*. Association for Computational Linguistics, Berlin, Germany, pages 54–61. <https://doi.org/10.18653/v1/W16-2009>.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. **A shared task on multimodal machine translation and crosslingual image description**. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 543–553. <https://doi.org/10.18653/v1/W16-2346>.
- Sanford B Steever. 2015. *The Dravidian Languages*. Routledge.
- Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybylski, and Signe Gilbro. 2014. **An overview of the european union's highly multilingual parallel corpora**. *Language Resources and Evaluation* 48(4):679–707. <https://doi.org/10.1007/s10579-014-9277-0>.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. **Sequence to sequence learning with neural networks**. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. MIT Press, Cambridge, MA, USA, NIPS'14, pages 3104–3112. <http://dl.acm.org/citation.cfm?id=2969033.2969173>.
- 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*. Association for Computational Linguistics, Jeju Island, Korea, pages 24–36. <https://www.aclweb.org/anthology/D12-1003>.
- Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.
- Jörg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, Borovets, Bulgaria, volume V, chapter V, pages 237–248.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*. volume 2012, pages 2214–2218.
- Jörg Tiedemann, Fabienne Cap, Jenna Kanerva, Filip Ginter, Sara Stymne, Robert Östling, and Marion Weller-Di Marco. 2016. **Phrase-based SMT for Finnish with more data, better models and alternative alignment and translation tools**. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Association for Computational Linguistics, Berlin, Germany, pages 391–398. <https://doi.org/10.18653/v1/W16-2326>.
- Jörg Tiedemann and Lars Nygaard. 2004. **The OPUS Corpus - Parallel and Free: <http://logos.uio.no/opus>**. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA). <http://www.aclweb.org/anthology/L04-1174>.
- Kristina Toutanova and Robert Moore. 2002. **Pronunciation modeling for improved spelling correction**. In *Proceedings of the 40th Annual Meeting of the Association for Computational*

## Bibliography

---

- Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 144–151. <https://doi.org/10.3115/1073083.1073109>.
- 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*. Association for Computational Linguistics, Montreal, Quebec, Canada, pages 1299–1306. <https://doi.org/10.3115/980691.980781>.
- Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. **Neural machine translation incorporating named entity**. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pages 3240–3250. <https://www.aclweb.org/anthology/C18-1274>.
- Masao Utiyama and Hitoshi Isahara. 2007. **A comparison of pivot methods for phrase-based statistical machine translation**. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Association for Computational Linguistics, Rochester, New York, pages 484–491. <https://www.aclweb.org/anthology/N07-1061>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need**. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., USA, NIPS’17, pages 6000–6010. <http://dl.acm.org/citation.cfm?id=3295222.3295349>.
- T. N. Vikram and Shalini R. Urs. 2007. *Development of Prototype Morphological Analyzer for the South Indian Language of Kannada*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 109–116.
- 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)*. Association for Computational Linguistics, Berlin, Germany, pages 247–257. <https://doi.org/10.18653/v1/P16-1024>.
- 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)*. Association for Computational Linguistics, Beijing, China, pages 719–725. <https://doi.org/10.3115/v1/P15-2118>.
- Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019a. **Learning deep transformer models for machine translation**. In *Proceedings of the 57th*

- Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 1810–1822. <https://doi.org/10.18653/v1/P19-1176>.
- Ye-Yi Wang and Alex Waibel. 1997. **Decoding algorithm in statistical machine translation**. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Madrid, Spain, pages 366–372. <https://doi.org/10.3115/976909.979664>.
- Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019b. **A compact and language-sensitive multilingual translation method**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 1213–1223. <https://doi.org/10.18653/v1/P19-1117>.
- Hua Wu and Haifeng Wang. 2007. **Pivot language approach for phrase-based statistical machine translation**. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 856–863. <https://www.aclweb.org/anthology/P07-1108>.
- Hua Wu and Haifeng Wang. 2009. **Revisiting pivot language approach for machine translation**. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, pages 154–162. <https://www.aclweb.org/anthology/P09-1018>.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. **Generalized data augmentation for low-resource translation**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 5786–5796. <https://doi.org/10.18653/v1/P19-1579>.
- Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2016. **A character-aware encoder for neural machine translation**. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 3063–3070. <https://www.aclweb.org/anthology/C16-1288>.
- Michael Yoder, Shruti Rijhwani, Carolyn Rosé, and Lori Levin. 2017. **Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages**. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, pages 73–82. <http://aclweb.org/anthology/W17-2911>.