# Neural and Multilingual Approaches to Machine Translation for Indian Languages and its Applications

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computer Science and Engineering** by Research*

by

Jerin Philip
201401071
jerin.philip@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
August 2020

International Institute of Information Technology
Hyderabad, India

# CERTIFICATE

It is certified that the work contained in this thesis, titled “Neural and Multilingual Approaches to Machine Translation for Indian Languages and its Applications” by Jerin Philip, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. C.V Jawahar

_____ _____

Date Co-Adviser: Dr. Vinay P. Namboodiri

To computational resources, and a terrific transition in risk-appetite over a short span of time.

# Acknowledgments

I am grateful for the free-hand to make choices and constant support provided by my parents without which made possible the pursuit of my research interests and hence this work.

This work is made possible by my advisors Prof. C.V. Jawahar and Prof. Vinay P. Namboodiri. They have guided me from the basics to where I am while wrapping up this work through times of failures, success and *uncertainty*[1]. I am grateful to Prof. C.V. Jawahar for his experienced guidance while taking my first steps as a research student and that I was given a perspective into not just publications, but diverse spheres of the academia. I believe I was provided with opportunities more than in a strict publication-only direction, which has contributed to an all-round personal and academic development. Prof. Jawahar has had a significant influence in my transition from a rigid set of programmer-like thinking to a research-ready mindset, where I am flexible and more comfortable with the existence of uncertainty that comes with working in frontiers of scientific knowledge. I express my sincere gratitude to Prof. Vinay, who continually provides insightful ideas and discussions, many of which made way into this thesis. A part of these I have managed to convert in this work, while a more extensive set of ideas and questions which I have not. I take this space to acknowledge the failures and those incomplete and ensure the ideas live-on for fruition in the future. I will especially remember the times Prof. Vinay has patiently instilled confidence to move forward in an otherwise initially confused and scared modus-operandi. The access both my advisors have provided me to the lab, its computational resources, insights into the academia have taught me much more than a skillset which is restrictive to research.

I thank the CVIT ecosystem - people, hardware and software for support and inspiration. I express my sincere gratitude to people who are both my friends and excellent researchers (or at least on the way there…) - Deepayan Das, Sangeeth Reddy, Aditya Arun, Avijit Dasgupta, Chris Andrew, Shyam Nandan Rai, Vamsi Muthireddy, Harish Krishna, Binu Jasim, Minesh Mathew, Praveen Krishnan and Pritish Mohapatra. Shashank Siripragada deserves a special mention, being a co-author in several of the publications which ended up in thesis, providing valuable contributions. I am grateful to the staff - Siva Kumar, Silar Shaik and Rohitha for making otherwise tedious administrative routines orders easier.

I thank Sandhu, Sid, Dorbala and Sahil for at times actively engaging in parts my life, pointless entertaining banter, dragging me to places where I otherwise would not usually have and enabling my caffeine addiction. I thank Minesh, Mithun, Praveen, Bhargu, Aquib and the many others part of the

[1] In chronological order, pun intended.

Kerala diaspora within IIIT who helped me while away, keep in touch with my roots in my home state - its food, movies and culture.

Much of the compilation and shoddy cut-paste job which rendered this thesis was done in free-time in weekends under the comfortable and chill weather of Grenoble, France and amidst a heat-wave in Toulouse, France. I thank the warm people of NAVER LABS Europe, especially my supervisors Alexandre Berard, Prof. Laurent Besacier and Matthias Galle. I thank Singhamaneni Phani Teja, who temporarily hosted me in Toulouse while being patient with my bicker about thesis, academia, research, IIIT and visa-troubles.

It makes sense to use this space also to mention and thank a certain someone whom I have so far met only once in life about a decade ago; but we somehow end up crossing paths in similar junctures in life. Our exciting conversations about shared interests have been a space of happiness and good feeling when things temporarily take a turn for the worse in the main work thread.

# Abstract

Neural Machine Translation (NMT), together with multilingual formulations have arisen as the de-facto standard in translating a sentence from a source language to a target language. However, unlike many western languages, the available resources like training data of parallel sentences or trained models which can be used to build and demonstrate applications in other domains are limited for the languages in the Indian subcontinent. This work takes a major step towards closing this gap.

In this work, we describe the development of state-of-the art translation solutions for 10 Indian languages and English. We do this in four parts described below:

1. Considering the Hindi-English language pair we successfully develop an NMT solution for a narrow-domain, demonstrating its application in translating cricket commentary.

2. Through heavy data augmentation, we extend the above to the general domain and build a state-of-the art MT system for Hindi-English language pair. Further, We extend to five more languages by taking advantage of multiway formulations.

3. We demonstrate the application of the NMT in contributing more resources to the already resource-scarce field, expanding to 11 langauges and its application in a multimodal task of translating a talking face to a target language with lip synchronization.

4. Next, we improve both data-situation and performance for machine translation in 11 Indian Languages iteratively to place our models in a standardized, comparable set of metrics setting up for future advances in the space to comprehensively evaluate and compare against.

# Contents


Chapter Page

1 Introduction . . . 1
1.1 Contributions . . . 2
1.2 Thesis Organization . . . 2

2 Background and Related Work . . . 3
2.1 Machine Translation . . . 3
2.2 Neural Machine Translation (NMT) . . . 4
2.2.1 Open Vocabulary . . . 4
2.2.2 Encoder-Decoder Architecture . . . 6
2.2.2.1 Encoder-Decoder with Attention . . . 7
2.2.3 Implementations of Encoder-Decoder Paradigm . . . 8
2.2.3.1 Recurrent Neural Networks . . . 8
2.2.3.2 Convolutional Sequence to Sequence Models . . . 9
2.2.3.3 Transformer Networks . . . 9
2.2.4 Multilingual NMT . . . 12
2.2.5 Backtranslation . . . 12
2.3 Machine Translation in Indian Languages . . . 13
2.3.1 Text Resources . . . 13
2.3.2 Linguistic Aspects . . . 14
2.3.3 Methods . . . 14
2.3.4 Evaluations and Benchmarks . . . 15

3 NMT for English-Hindi . . . 16
3.1 Translating a Narrow Domain . . . 16
3.1.1 Introduction . . . 16
3.1.2 Related Work . . . 17
3.1.3 Motivation . . . 18
3.1.4 Neural Machine Translation . . . 20
3.1.4.1 Encoder - Decoder . . . 20
3.1.4.2 Attention . . . 21
3.1.5 Experimental Setup . . . 21
3.1.5.1 Dataset . . . 21
3.1.5.2 Architecture and Framework . . . 22
3.1.5.3 Methods . . . 22
3.1.5.4 Evaluations . . . 22

3.1.6 Results and Discussions . . . 23
3.1.6.1 Quantitative Results . . . 23
3.1.6.2 Qualitative Results . . . 23
3.1.7 Cricket Commentary . . . 26
3.1.8 Conclusions and Future Work . . . 27
3.2 Translating General Domain . . . 28
3.2.1 Components . . . 28
3.2.1.1 Tokenization . . . 28
3.2.1.2 Convolutional Sequence to Sequence Learning . . . 28
3.2.2 Experimental Setup . . . 29
3.2.2.1 Dataset . . . 29
3.2.2.2 Data Processing . . . 29
3.2.2.3 Training . . . 30
3.2.2.4 Evaluations . . . 31
3.2.2.5 Discussions . . . 31
3.2.3 Transformer Experiments . . . 31
3.3 Conclusion . . . 32

4 Multilingual Approaches for Low Resource languages . . . 33
4.1 Datasets . . . 33
4.2 Multilingual Neural Machine Translation . . . 34
4.2.1 Evaluation Protocol . . . 36
4.2.2 Results . . . 36
4.2.2.1 On IITB-en-hi and WAT-ILMPC . . . 36
4.2.2.2 Multilingual Results . . . 37
4.2.3 Qualitative Examples and Discussions . . . 38
4.3 Discussions . . . 38
4.3.1 More on WAT-ILMPC . . . 38
4.3.2 Mann Ki Baat . . . 42
4.4 Revisiting translation in narrow domain . . . 46
4.5 Conclusions . . . 46

5 Applications . . . 48
5.1 Enhancing translation resources for Indian Languages . . . 48
5.1.1 Motivation . . . 48
5.1.2 Aligned Parallel Corpora from Web . . . 50
5.1.3 Dataset Description and Statistics . . . 53
5.1.3.1 Press Information Bureau Corpus . . . 53
5.1.3.2 Mann Ki Baat . . . 54
5.1.3.3 Characterizations and Comparisons . . . 55
5.1.4 Discussions . . . 55
5.1.4.1 Alignment Quality . . . 55
5.1.4.2 Translation Quality . . . 59
5.1.4.3 PIB for NMT . . . 60
5.1.5 Conclusion and Future Work . . . 62
5.1.6 Release Information . . . 64

5.2 Automatic Face to Face Translation . . . 64
5.2.1 Translating to target language $L_B$ . . . 65

6 Revisiting Low Resource Status of Indian Languages in Machine Translation . . . 67
6.1 Introduction . . . 67
6.2 Iterative Alignment for PIB . . . 70
6.2.1 Data Sources . . . 70
6.2.2 Iterative Alignment . . . 71
6.2.2.1 Text Processing . . . 72
6.2.2.2 Alignment Algorithms . . . 72
6.2.2.3 Multilingual Neural Machine Translation (MNMT) Model . . . 73
6.2.2.4 Iterations . . . 74
6.2.3 Discussions . . . 74
6.3 Stronger and Repeatable Baselines . . . 78
6.3.1 On Repeatability of Objective Evaluations . . . 78
6.3.1.1 Test Sets . . . 78
6.3.1.2 Comparable reports of BLEU Scores . . . 78
6.3.2 Results and Discussions . . . 79
6.3.3 Comparison with Previous Works . . . 80
6.3.4 Summary: Stronger Baselines and Public Benchmarks . . . 82
6.4 Discussions and Future Directions . . . 83

7 Conclusion . . . 84

Bibliography . . . 86

# List of Figures


Figure Page

2.1 ConvS2S architecture from [27]. . . . 10
2.2 Transformer architecture from [98]. . . . 11
2.3 Growth of Wikipedia pages in languages . . . . 14

3.1 Corpus growth . . . . 19
3.2 Scenes from a cricket match overlayed with subtitles. Possible application of this work. English overlays are shown in the first row, and the corresponding translated Hindi commentary below. . . . 26

5.1 The Face-to-Face translation pipeline. NMT is a critical component where the language-crossing happens. . . . 64

6.1 Iterative alignment pipeline used for expanding the corpus for Indian languages. We observe that (i) A better MNMT model leads to better alignment and larger corpus (ii) Larger corpus leads to better MNMT model. We iterate until no further improvement is observed. The dashed lines indicate application of the trained MNMT model. *pp* stands for an arbitrary pivot language. . . . 69
6.2 Figure illustrating number of articles on the Y-axis obtained at a given threshold on X-axis. We maintain a constant threshold (dotted line) of *0.51* across model iterations M2M-0, M2EN-1 and M2EN-2. In case of Marathi, when compared to M2M-0 we acquire an additional 2.5K article pairs using M2EN-1 and 0.8K more using M2EN-2. From the graph, we observe saturation in articles pairs after iteration 2 indicating a point of diminishing returns. . . . 75
6.3 Retrieval scores of *Gujarati*. Left bar chart indicates retrieval scores in case of model M2M-1 and pivot language *hi*. Right chart indicates scores in case of M2EN-1 and pivot *en*. . . . 76

# List of Tables


Table Page

3.1 Failure cases of a generic translation system. . . . 19
3.2 Details of the corpus used. . . . 22
3.3 Comparison our results with the one provided by generic translation system. BLEU, METEOR, ROUGE_L, WER are reported in percentages. . . . 23
3.4 Qualitative Results and Comparisons: English to Hindi. g-pred is the prediction of a generic system, while pred is our prediction. . . . 24
3.5 Qualitative Results and Comparisons . . . 25
3.6 Descriptions of the corpora used, IITB train† is a filtered version of the IITB train corpus. 29
3.7 Quantitative results of translating English to Hindi and vice versa. . . . 30
3.8 Transformer-Base vs ConvS2S on National News + IITB corpus, for English to Hindi direction. . . . 32

4.1 WAT Corpus Details . . . 33
4.2 Five example sentences in English and their translations in Hindi. Readers may note the (i) length (ii) complexity and (iii) diversity among models of the sentences . . . 35
4.3 BLEU scores on WAT-ILMPC (wat-*xx*) and IITB-hi-en (iitb-*xx*) test sets. Directions are indicated in the first column. Our model is the best in English-to-Hindi, in the top 3 in Hindi to English in IITB-hi-en. In addition to this, it performs well on WAT-ILMPC evaluation metrics. . . . 36
4.4 BLEU scores over translation pairs in Mann Ki Baat multilingual test-set. The complexity of sentences are high, with long and rich sentences of the kind included in ministerial speeches. . . . 37
4.5 Success Cases from WAT test-set. . . . 39
4.6 Failure Cases from WAT test-set . . . 40
4.7 Qualitative sample from WAT test-dataset. Missing tokens are colored in red. Few samples which are treated as gold aren't as good. . . . 41
4.8 BLEU scores of *IL-MULTI+ft* . . . 43
4.9 BLEU scores of *IL-MULTI* . . . 43
4.10 BLEU scores of *IL-MULTI+bt* . . . 43
4.11 Qualitative samples from Mann Ki Baat Dataset. Source is one large row. Translations to different languages are shown in two columns - first one is our prediction and second one the respective aligned ground truth. . . . 44
4.12 Longer sentences from Mann Ki Baat set. . . . 45

4.13 BLEU scores of different takes on building a domain specific translation system, for cricket. Cold-start denotes model trained on cricket data alone. Warm start begin with *IL-MULTI* - trained on the general dataset. This model is further fine-tuned to cricket data. . . . 46

5.1 Detailed statistics of PIB and Mann Ki Baat. Number of sentences are reported after segmentation of articles. Aligned-en indicates number of sentences aligned to English across each language. Filtered indicates size after filtering aligned sentences through our pipeline. We also report vocabulary compiled from filtered sentences across languages. Details on nature of PIB and Mann Ki Baat corpus in sections 5.1.3.1 and 5.1.3.2 respectively. †Punjabi(pa) is unavailable in Mann Ki Baat. . . . 51

5.2 Training dataset used for multilingual model. xx-yy indicates parallel sentences aligned across multiple languages. . . . 52

5.3 Success cases of Alignment for sentence pairs from PIB and Mann Ki Baat. 1 and 3 are shorter sentences but relatively hard to align to English due to presence of sentence delimiters on the English side. 2 and 4 have longer Hindi sentence containing two segments delimited by *Purnavirama* (Hindi end of sentence marker) symbol. We observe that *BLEUAlign* merges these segments to obtain a higher BLEU score, aligning to a longer Hindi (1&3) and English (2&4) sentence which conveys same meaning as the source. . . . 56

5.4 Failure cases of Alignment for sentence pairs from PIB and Mann Ki Baat. 1 represents a hard case to align, where the content is written elaborately in source and poorly in target language. 2 contains English sentence with punctuation, misaligned due to segmentation. 3 represents a case of partial alignment due to segmentation of the English sentence. . . . 57

5.5 Multilingual datasets available for Indian languages. . . . 57

5.6 Multilingual shared content across language pairs for PIB (Blue) and Mann Ki Baat (Red). Rows and columns indicate language pairs. Upper triangle (Blue) indicates PIB across Languages. Lower triangle (Red) indicates Mann Ki Baat. The intensity of the cell color is proportional to size of the sentences aligned pairs. † Mann Ki Baat does not contain Punjabi. . . . 58

5.7 Assessment of sentence alignment quality. . . . 59

5.8 BLEU scores of model trained on unaugmented data on multilingual test sets - test-split sampled from ILCI from the unaugmented model. Rows correspond to source languages and columns target languages. . . . 60

5.9 BLEU scores of model trained on unaugmented data on multilingual test sets - test-split sampled from Mann-Ki Baat obtained from the unaugmented model. Rows correspond to source languages and columns target languages. . . . 61

5.10 Cross domain inference. Rows indicate the corpus on which model is trained and columns the test-corpus. BLEU scores are reported for en-hi. All data is listed in Table 5.2 . . . 62

5.11 Improvements (differences in BLEU) in directions by augmenting with PIB vs un-augmented with all remaining data. Reds indicate drop in BLEU scores and blues indicate improvements. We see improvements overall by using PIB as augmented data for training. . . . 63

5.12 NMT Evaluation Scores. . . . 65

6.1 Publicly available corpuses for Indian languages. The last group of rows were not used for training. CVIT Mann Ki Baat is used for evaluation purposes only and has overlap with PMIndia Corpus. All other sources are used for training the multilingual model. xx-yy indicates parallel sentences aligned across multiple languages. Last row is the proposed corpus. . . . . 71
6.2 Incremental improvements in Accuracy, *xx→en* BLEU scores on *CVIT Mann Ki Baat* and Corpus size. We observe increments in retrieval accuracies consistent with increase in BLEU scores. We increment the initial version of the PIB corpus with an additional *201K* sentences aligned to English. (*CVIT Mann Ki Baat* does not contain Punjabi.) . . 73
6.3 Multilingual shared content across language pairs for CVIT-PIBv0.2. Rows and columns indicate language pairs. The highlights are proportional to the change after the iterative alignment process, reds indicating decrease and blues indicating increases in corpora sizes compared to the previous release v0.0. Evident from the table, we notice major increments in Marathi, Oriya and other languages. Tamil and Hindi which we had enough to be considered mid-to-high resource gain significant number as well. The maximum decrease is -283 for Telugu, which is negligible compared to the improvements of the order of ten-thousands in many language-pairs. . . . . 77
6.4 We report BLEU scores on available publicly available benchmark tasks for Indian Languages. The results on these benchmarks often have models that are specially tuned for various language pairs. We do observe that we obtain state of the art results on 3 of the language pairs and are competitive to other works that are more specific in most cases. This is despite not being specially tuned for these settings. OEC stands for OdiEnCorp, W19, W20 stands for WMT19 and WMT20. [1] stands for test-split and [2] stands for dev-split respectively. . . . . 79
6.5 Comparison with publicly available baselines for English to Hindi and vice versa. . . . 81
6.6 BLEU scores of M2M-3 model on multilingual test set Mann-Ki Baat. Rows correspond to source languages and columns target languages. The colors indicate improvement (blue) or degradation (red) in comparison to M2M-1 [93]. We observe cumulative increment of +257 BLEU across all language pairs and a median increment of +1.25. The cumulative changes in translating to or from a given language in comparison to M2M-1 are provided under $\Delta$ header. It can be observed that related languages end up with higher BLEU scores without having to add the prior in the model formulation - e.g (hi, gu), (ur, gu), (ur, hi). Closely behind, there is (mr, hi) ahead of other language pairs. While there seems to be a decrease in *en→xx* between M2M-1 and M2M-3 on Table 6.4, the overall improvements here indicate that it is not the general case. . . . . 82

# Chapter *1*

# Introduction

Machine Translation (MT) [11] is the umbrella term used for methods obtaining translation from a source language to a target language. The past decade has seen massive improvements in the solutions to this problem, driven by developments in deep-learning [55]. This class of solutions, called Neural Machine Translation (NMT) has since been applied to enable near-human level translations over several languages.

NMT has demonstrated effectiveness in the supervised learning setting, where labelled data is available and in this case is a parallel-corpus. A parallel-corpus is a sentence-level aligned collection of sentences between the source and target language. The success of NMT methods has been reported on high-resource languages [6]. However, further improvements have propelled the success of NMT in low-resource settings. One set of improvements are brought about by data-augmentation through synthetic parallel-corpora [86, 24]. Transfer Learning approaches [102] have an orthogonal set of improvements in low-resource languages.

Effective methods of translation between languages which had no data aligned between them were brought about previously by translating them through a common "pivot" language, which in most cases is English due to the availability of data. Unlike this explicit pivoting, by being able to implicitly pivot during training through the English data brought about considerable improvements in the zero-shot setting. In the zero-shot setting in the context of NMT, a system can translate between language pairs unseen during the time of training. These multilingual-formulations proposed by [36] enabled massive boosts to low resource languages. In addition to this, the sharing of parameters among language-directions additionally allowed a significant reduction in model sizes.

So far, there have been limited explorations of approaches using NMT in Indian Languages. Further, the lack of availability of parallel-corpora among languages put many languages in the low-resource category. The above two also lead to fewer attempts as NMT for general-purpose translation requires a large volume of training data for supervision.

## 1.1 Contributions

In this work, we pursue NMT approaches to Indian Languages. The contributions of this work are as follows:

1. We experiment with English-Hindi and use data-augmentation approaches to build state of the art systems among the language-pairs.
2. Further, using multilingual formulations, we demonstrate the success of NMT models in low resource Indian languages.
3. Using the above systems, we create a multiparallel corpora which is a dataset of parallel sentences aligned across 11 languages, to further research in this direction.

## 1.2 Thesis Organization

The thesis is organized as follows: **Chapter 2** lays out the description of the components and literature related to this work. **Chapter 3** describes the creation of an NMT system which is currently state-of-the-art in English-Hindi languages. In **Chapter 4**, we use multilingual NMT approaches to improve the translation quality in language directions involving English and 7 Indian languages. Further in **Chapter 5**, we describe applications of the above NMT model in (1) increase parallel corpus for machine translation using our NMT systems and contributes a dataset aligned across 11 languages, (2) to a first-attempt at pipelining end-to-end face-to-face translation. Finally in **Chapter 6**, we describe our approach take on improving the data-situation and machine translation model performance alternatively optimizing for both. We conclude our findings in **Chapter 7** and discuss potential future directions.

# Chapter *2*

# Background and Related Work

## 2.1 Machine Translation

Early solutions took the form of rule-based systems where rules were programmed in by a human, termed rule-based machine translation (RBMT). With advances in statistical methods, using data to learn these rules and to resolve ambiguity in rules through context has been attempted by a class of methods under the umbrella of Statistical Machine Translation (SMT). Another class of solutions proposed transductive prediction of the target sentence, from several examples, called example-based MT (EBMT).

A statistical MT model uses the following formulation for a source sequence $\mathbf{x}$ and a target sequence $\mathbf{y}$:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \overbrace{P(\mathbf{x}|\mathbf{y})}^{\text{Translation Model}} \cdot \underbrace{P(\mathbf{y})}_{\text{Language Model}} \tag{2.1}$$

The probabilities $P(\mathbf{x}|\mathbf{y})$ and $P(\mathbf{y})$ is estimated using frequencies of phrase or word-units. The collection of phrases is restricted to a few words to make the computation of these probabilities feasible, decomposing them into products of factors. The computations of frequency tables etc. were parallelized over multiple CPU cores and computers providing the earliest usable translation systems for the public [42]. SMT brought about significant improvements to automatic translations to the point it was deployed in popular online services like Google Translate.

Parallel to the success of neural networks in image recognition, speech recognition, etc., deep neural networks (DNNs) have found widespread use to evolve as the de-facto method at the time pursuing this work. The class of solutions using deep neural networks to learn from data a translation model is widely known as Neural Machine Translation (NMT) which is used extensively in this work. NMT is a different paradigm of GPU heavy approaches involving neural networks and learns by backpropagation, unlike the frequency-based procedures in SMT. We discuss elaborately the components that make up modern NMT ahead.

## 2.2 Neural Machine Translation (NMT)

This section introduces and elaborately describes the building blocks for the NMT approaches used in this thesis. First, machine translation is cast as a sequence-to-sequence learning problem. With such a formulation, methods to decompose sentences as sequences constituted by meaningful units are required to make implementation feasible. Open-vocabulary methods used in this work to achieve the same are discussed. Two variations of neural network architectures used in this work are described in detail.

### 2.2.1 Open Vocabulary

A fundamental step in training models in deep neural networks with text is the conversion of the discrete text sequence into a sequence of meaningful vector representations. Early approaches [96, 6] restricted NMT to the most frequent words in the vocabulary generated from the source-text available for a language. With this setting, an embedding-matrix $E$ converts the words represented by the one-hot vectors corresponding to the word to a smaller dimension real vector.

However, the above approach of limiting words is suboptimal, as (1) it does not take into account redundancies like inflections, and (2) one cannot cover the whole vocabulary reading to many unknowns and rare-words rendered unable to translate. Methods to circumvent this try to maximize language coverage with a given vocabulary limit, falling under the class of open-vocabulary methods. Two popular approaches to achieve open vocabulary for in natural language applications are Byte Pair Encoding (BPE) [85] and SentencePiece [45]. The objective is to maximize coverage in the corpus given minimum "subword" units - where SentencePiece uses an entropy-based formulation while BPE uses a dictionary-based formulation. A subword here is a statistically motivated unit of language. There are linguistically motivated units which could be used alternatively. But given the number of languages covered in this work and the requirement that our system generalizes to the many languages involved, statistically motivated units are used.

#### Byte Pair Encoding

The input and output spaces are made up of classes which correspond to the words in source and target vocabulary, respectively. For the models to work feasibly, a finite number of classes are necessary. Often in many languages with flexible word boundary, using space-separated words as classes end up with an exponentially growing vocabulary due to inflections and agglutinations. Byte Pair Encoding (BPE) is a technique from compression literature, which recursively replaces most frequently occurring pairs of symbols with a new symbol of constant space. Applying BPE to a monolingual corpus has been shown to produce improvements in translation [85], especially when treatment as word units leads to a lot of out of vocabulary words (OOV). In contrast, treatment as subword units helps reduce vocabulary. Our pipeline employs BPE for its benefits mentioned above.

**Algorithm 1** Learn BPE operations: as shown by Sennrich et al. [85]

```
import re, collections

def get_stats(vocab):
  pairs = collections.defaultdict(int)
  for word, freq in vocab.items():
    symbols = word.split()
    for i in range(len(symbols)-1):
      pairs[symbols[i],symbols[i+1]] += freq
  return pairs

def merge_vocab(pair, v_in):
  v_out = {}
  bigram = re.escape(' '.join(pair))
  p = re.compile(r'(?<!\S)' + bigram + r'(?!\S)')
  for word in v_in:
    w_out = p.sub(''.join(pair), word)
    v_out[w_out] = v_in[word]
  return v_out

vocab = {'l o w </w>' : 5, 'l o w e r </w>' : 2,
         'n e w e s t </w>':6, 'w i d e s t </w>':3}
num_merges = 10
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
  print(best)
```

**SentencePiece**

Kudo [45] presents an alternate formulation, where the objective is to increase the likelihood of the corpus by varying a set of tokens.

This work uses specifically the segmentation model involving a unigram language-model, where a unigram language model is trained to minimize the following loss:

$$\mathcal{L} = \sum_{s=1}^{|D|} \log P(X^{(s)}) = \sum_{s=1}^{|D|} \log \left( \sum_{x \in S(X^{(s)})} P(x) \right) \tag{2.2}$$

$D$ is the monolingual corpus in the particular language for which the subword vocabulary is to be created. Methods prevalent for western languages use a shared vocabulary across several languages, but however in this work due to the several different scripts involved in Indian languages and imbalance in training data amongst languages, we use separate unigram models for each language.

Since we do not know the subwords (thus the vocabulary) at the beginning of the process, an expectation-maximization procedure is used to reach a satisfactory subword vocabulary. This is accompanied by an initialization, which starts with the initial vocabulary being generated from words separated by the space delimiter. The unigram model is then fit with the vocabulary (expectation-step). In each update step of the vocabulary, the likelihood of the corpus by alternate subwords is evaluated, excluding a given subword. The subwords among those in vocabulary, which lead to the smallest decrease in loss are discarded (maximization-step) and the procedure repeats until it reaches the desired number of subwords in vocabulary.

### 2.2.2 Encoder-Decoder Architecture

Sequence to sequence models (seq2seq) with an encoder-decoder architecture are widely prevalent among NMT approaches [96, 6, 58, 98]. We also employ a similar architecture for translating text in one language to another. We now proceed to formally describe the parts relevant to this work below.

Consider a source sequence $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ and a target sequence $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, where $x_i, y_j \in V$. Our problem is to translate the sequence $\mathbf{x}$ to $\mathbf{y}$. Our translation model has an encoder ($f_\theta(\cdot)$) followed by a decoder. The encoder consumes the source side sequence $\mathbf{x}$ to learn representation $\mathbf{c}$. The auto-regressive decoder learns to maximize log-likelihood of the target sequence modelled by the conditional probability distribution given below. For a given time-step in the decoding process $t$, the generation process for the token at $t$ is conditioned on $\mathbf{c}$ and the tokens generated until $t-1$.

The formulation now can be described as follows:

$$\mathbf{c} = f(\mathbf{x}; \theta_E) \tag{2.3}$$

$$L(\theta_E, \theta_D) = \sum_{t=1}^{t=T} \log p(y_t | y_{t'<t}, \mathbf{c}; \theta_\mathbf{D}) \tag{2.4}$$

$$\tag{2.5}$$

All $(\mathbf{x}, \mathbf{y})$ pairs we have are mini-batched to fit onto resources available. The learnable parameter $\theta$ is optimized through mini-batch gradient descent of the objective. In this work, updates, steps or mini-batches are used synonymously.

**Decoding**

A model trained with the objective and training procedure above now predicts the sequence at each step in an autoregressive manner. The simplest method to generate a target sequence during testing is called **Greedy Decoding** and can be described as follows:

$$g(t) = \arg\max_{y_t \in V} \log p(y_t | y_{t'<t}, \mathbf{c}; \theta) \tag{2.6}$$

The sequences are in practice wrapped in a control-token which indicates a start-of-sequence or end-of-sequence while training. This allows starting the decoding process with the start-of-sequence token and autoregressively doing the above generates the target sequence and this process stops only once the end-of-sequence is generated.

However, the above formulation is plagued by something called the *exposure bias* at the decoder - error at an arbitrary $t$-th decoding time-step is propagated and likely amplified until $T$, the end of the prediction sequence. A generalization, **beam search** helps to reduce the issue to some extent. At each time-step of prediction, the $k$-best hypotheses are maintained, where $k$ corresponds to the beam width. Beam search also enables producing multiple candidate suggestions for translation of a single source - which are ranked by probabilities from the decoder. In our experiments, we only take the best translation, while the beam in place helps alleviate error accumulation.

#### 2.2.2.1 Encoder-Decoder with Attention

Sutskever et al. [96] proposed a successful sequence to sequence model which could perform translation task which was constituted by Long Short Term Memory (LSTMs). Bahdanau et al. [6] improves upon Sutskever et al. [96] by introducing attention, which boosted performance further. Attention enhanced the decoder with the new ability to use more context than a single vector and the ability to attend to several source sequences. Luong et al. [58]'s methods further study attention-based models. The modification brings about the following change in the above formulation:

$$\mathbf{h}_{1:T} = f(\mathbf{x}; \theta_E) \tag{2.7}$$

$$\mathbf{a}_t = \text{attn}(\mathbf{y}_t, \mathbf{h}_{1:T}; \theta_A) \tag{2.8}$$

$$\mathbf{c}_t = \sum_{i=1}^{T} a_{t,i} \mathbf{h}_i \tag{2.9}$$

$$L(\theta_E, \theta_A, \theta_D) = \sum_{t=1}^{t=T} \log p_\theta(y_t | y_{t'<t}, \mathbf{c_t}; \theta_D) \tag{2.10}$$

Gehring et al. [27] proposes a better model which involves the use of convolutional neural network (CNN) components instead of RNN units. The advantage over RNN for the ConvS2S model is that operations can now be parallelized. Vaswani et al. [98] takes attention from [6, 58] and uses it as a building block to create a new class of models called Transformer, which improves on performance in the translation task and incorporates parallelizability in the pipeline. This work uses both the convolutional and Transformer architectures, which are further explained below.

### 2.2.3 Implementations of Encoder-Decoder Paradigm

This work uses three implementations of the encoder-decoder paradigm on using Recurrent Neural Networks (RNNs) with attention [6, 58], convolutional sequence to sequence model (ConvS2S) [27] and Transformer-networks [98]. Below, we layout the description of these architectures and how they are realized.

#### 2.2.3.1 Recurrent Neural Networks

One popular choice of architecture for encoder and decoder has been the class of RNNs - usually Long Short Term Memory (LSTM) or Gated Recurrent Units (GRU) for learning longer dependencies. The encoder when realized using an RNN, where the hidden state at a given time-step is denoted by $\overline{h}_t$ and inputs and outputs denoted by $x_t$ and $z_t$, produces a context vector $\mathbf{c}$ as follows:

$$z_t, \overline{h}_t = \text{Encoder-RNN}(x_t, \overline{h}_{t-1})$$
$$\mathbf{c} = \overline{h}_{T'}$$

The decoder modelled by an RNN now generates the target side representation $\mathbf{y}$ of the source sequence $\mathbf{x}$ as follows:

$$h_0 = \mathbf{c}$$
$$\hat{y}_t, h_t = \text{Decoder-RNN}(\hat{y}_{t-1}, h_{t-1})$$
$$y_t = \text{softmax}(\hat{y}_t)$$

The early models used while pursuing this work use an implementation using stacked LSTMs for encoder and decoder.

**Attention** Luong et al. [58] discusses three ways of implementing attention. Our work uses the variant general attention, which accomplishes attending over input as follows. First, compute a score for each of the encoder input $\overline{h}_s$ with respect to hidden state at the current time-step $h_t$.

$$\begin{aligned} \text{score}(h_t, \overline{h}_s) &= h_t^T \mathbf{W_a} \overline{h}_s \\ \hat{\boldsymbol{a}}_t(s) &= \text{score}(h_t, \overline{h}_s) \end{aligned}$$

These can be converted to probability weights of attending to each of the encoder input, by applying softmax. Using the weights, a context vector **c** is derived as the weighted average of the source input states.

$$\begin{aligned} \boldsymbol{a}_t &= \text{softmax}(\hat{\boldsymbol{a}}_t) \\ \mathbf{c}_t &= \sum_s \boldsymbol{a}_t(s) \overline{h}_s \end{aligned}$$

#### 2.2.3.2 Convolutional Sequence to Sequence Models

A Gated Linear Unit (GLU) is the building block in a convolutional sequence to sequence model. Gehring et al. [27] demonstrates the first use of GLUs in implementing a sequence to sequence model.

With convolutional kernels parameterized by $W \in \mathbb{R}^{2d \times kd}$, a single layer takes an input $X \in \mathbb{R}^{kd}$, where $d$ is the dimension of the embeddings and $k$ entries present in the sequence, one can obtain $Y = [A, B] \in \mathbf{R}^{2d}$. A layer further performs the following operation with $[A, B]$ to obtain gating similar to RNNs.

$$v[A, B] = A \otimes \sigma(B)$$

Stacks of these layers are used in place of the RNNs to create the encoder and decoder. To enable deeper networks, skip-connections are added.

#### 2.2.3.3 Transformer Networks

Transformer architecture, proposed by Vaswani et al. [98] relies on self-attention alone. The construct of attention previously described by [6, 58] is generalized into the following. Transformer architectures powers most neural machine translation state-of-the-art performances. Transformer introduces two variants of attention - Scaled Dot Product and Multi-Head Attentions.

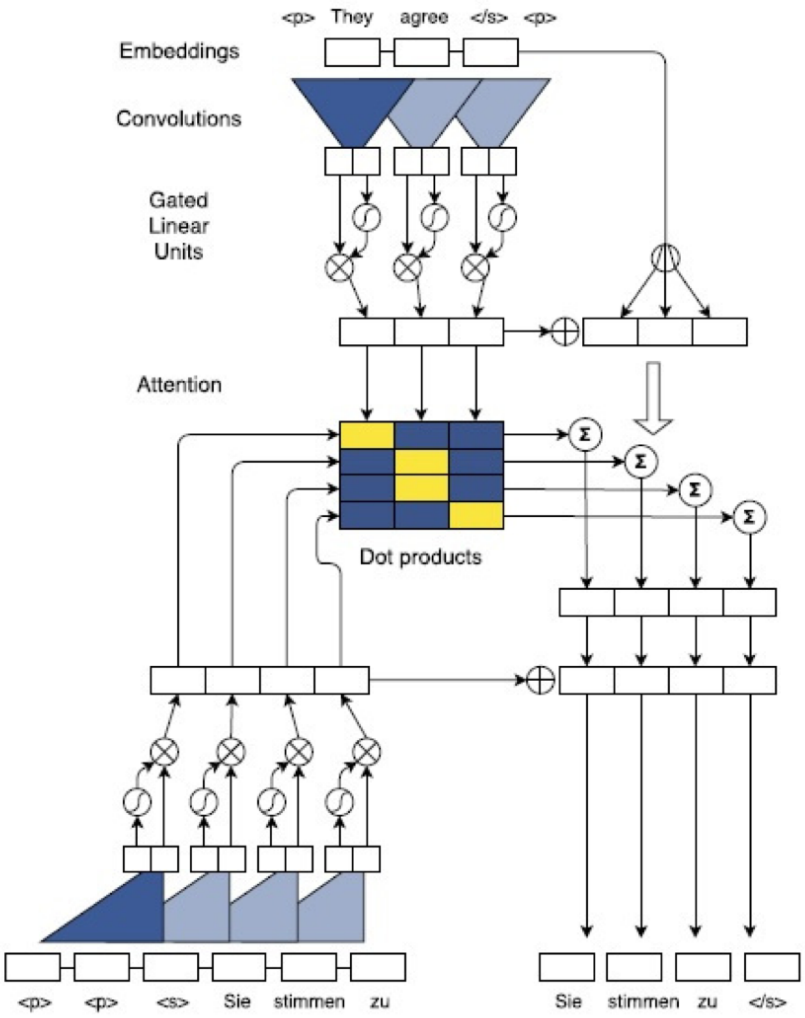


**Figure 2.1** ConvS2S architecture from [27].

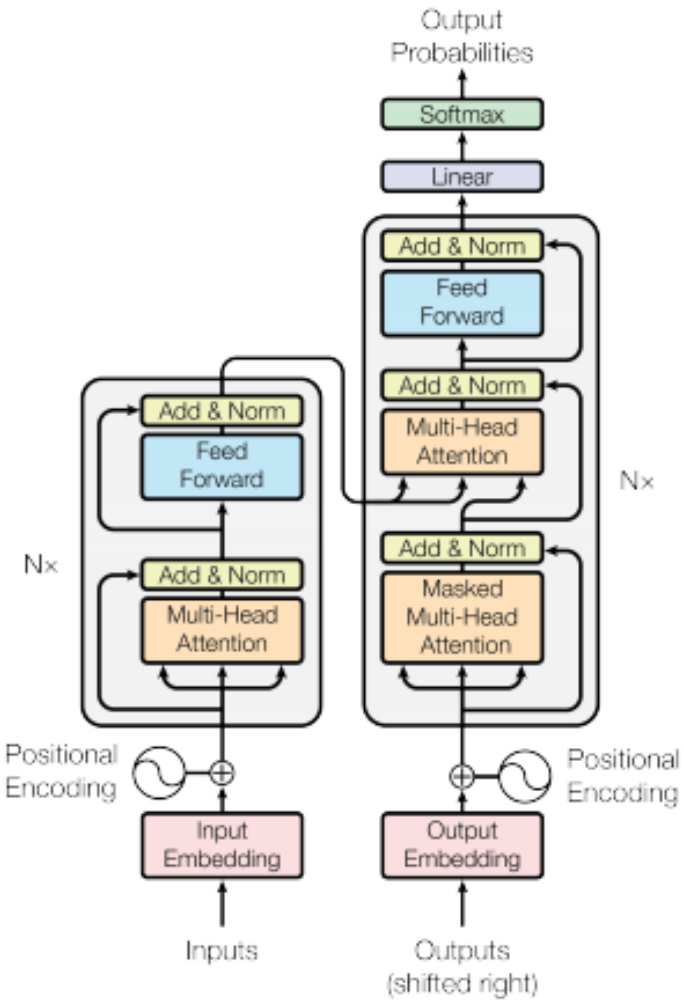


**Figure 2.2** Transformer architecture from [98].

Scaled Dot Product attention is simpler, but in practice Multi-Head Attentions are used due to their superior ability to jointly attend to information from different representational subspaces. $Q, K, V$ in the equation represents weight matrices which transform the queries, keys and values respectively.

Scaled dot product attention is described by the equations below:

$$\text{attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.11}$$

Multi-Head Attention is described by the following equations:

$$\text{MultiHeadAttention}(Q, K, V) = \text{concat}(\text{head}_1, \text{head}_2, \ldots \text{head}_n) \tag{2.12}$$

$$\text{where head}_i = \text{attn}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.13}$$

The encoder uses the embeddings of the source sequences, transformed by the corresponding weight parameters for queries, keys and values, thus giving it the name "self-attention". Over several stacks of such transformations, a refined sequence of encoder representations at each time-step is generated. The decoder uses the encoder-representations for keys, decoder-embeddings transformed for the queries and decoder-embeddings again for the values.

### 2.2.4 Multilingual NMT

In this work, we use the term Multilingual NMT to refer to models with parameters shared across several languages involved. Alternatively, this has been used in literature as "multiway" neural machine translation as well. This work uses the control-token based methods introduced by Johnson et al. [36] to implement multilingual NMT. In addition to the usual-control tokens (start-of-sequence, end-of-sequence) additional control tokens indicating the language to translate to are passed along with the input sequence. Now with this added control token in the training data, several language-directions for which data is available is fed to train a single-large model. This enables sharing of parameters of the encoder, decoder and attention across all languages.

The multilingual formulations in Johnson et al. [36] come with the added benefit of zero-shot capability amongst language-directions unseen during training. For example, if the training data were to have only English-centric parallel data, but several languages aligned with English - the multilingual model is able to translate amongst the directions involving languages that are not English. However in our case, while we merit from the zero-shot capability since we have seen data in all-directions the setting can be considered "few-shot", rather than "zero-shot".

### 2.2.5 Backtranslation

Backtranslation is a widely tried and tested data augmentation method, proposed for aiding NMT in languages low on parallel resources using available monolingual data by Sennrich et al. [86]. The method

works by first training a model in the low to high resource direction followed by using this model on monolingual data. The process provides more authentic sentences in the resource-scarce language and close approximation of its translation in the high resource language. It has been empirically shown that synthetic data alone generated through backtranslation can attain up to 83% of the performance using proper bitext [24]. This work uses backtranslation to demonstrate improvement in performance in several low-resource languages.

## 2.3 Machine Translation in Indian Languages

Machine translation systems could be built for a specific language pair or for multiple pairs simultaneously. We take the latter route in this work. This choice is dominated by many practical advantages in managing the limited available resources. We also hypothesize that the linguistic similarities across many Indian languages could help each other in this manner, though a systematic study on this is outside the scope of this work. Our work is in a multilingual setting wherein one system translates in all directions. We discuss challenges specific to the context of a few Indian languages.

### 2.3.1 Text Resources

Most Indian language pairs are resource-scarce in terms of sentence aligned parallel bi-text. More pairs and experts are likely to be available in pivoting through English to learn a multilingual translation model. Increasing reach of Internet access in the subcontinent is, however, changing these, as observable from Figure 2.3, which presents the systematic increase in the Wikipedia pages in Indian languages.

Hindi-English (hi-en) can be labelled a relatively high resource pair, after continued efforts to compile and increase resources over the past decade. The IIT-Bombay Hindi English Parallel Corpus (IITB-hi-en) [52] is the largest general domain dataset available for use in MT. In addition to a 1.5 Million parallel data, IITB-hi-en also offers nearly 19.2M monolingual Hindi text, which can be used to further enhance the performance.

The Indian Languages Corpora Initiative (ILCI) Corpus [35] is another major initiative, which offers corpus aligned in many Indian languages. The ILCI corpus we used here include parallel data specific to the domains of tourism and health across seven languages and 11 languages respectively in works that followed one another. Indic Languages Multilingual Parallel Corpus (WAT-ILMPC)[1] is a compilation of parallel data of subtitles, from OPUS[2], which contains parallel pairs of subtitles between languages of the subcontinent and English. We also make use of this.

In addition to the above, we attempt to use translations of religious texts like the bible through our own collection (Telugu) and [71], which includes the same in Oriya.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html

[2] http://opus.nlpl.eu/

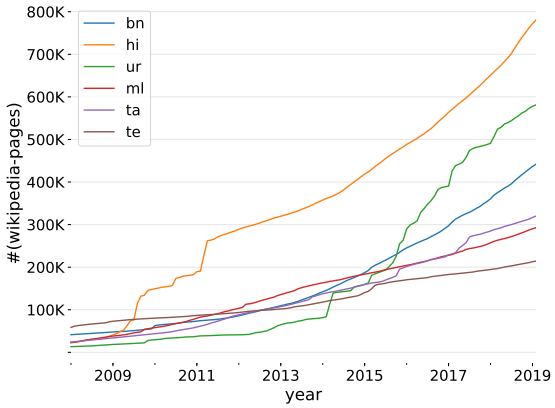


**Figure 2.3** Growth of Wikipedia pages in languages

### 2.3.2 Linguistic Aspects

Khan et al. [38] briefly covers linguistic aspects of several languages in the country. Indian languages are generally morphologically rich. Free word ordering in many languages lead to multiple ways of representing the content on the target side for some content in the source side. Dravidian languages face further challenges with heavy agglutinative and inflective characteristics. Recent developments in the use of open vocabulary [45, 85] have found success for many languages without large deterioration to translation quality. The same principles also help us in practically circumventing the challenges due to word morphology.

Our multilingual setting involves language pairs where translation is an "ill-posed" problem. For example, a single word "you" in English could translate to "tum", "me" and "aap" in Hindi, and valid sentences exist where correct translation can not be identified even with paragraph-level contexts. Addressing these are left out of the scope of this paper.

### 2.3.3 Methods

Most of the past attempts look at the problem of machine translation in Indian setting as a pair-wise translation problem. That is, the scope is restricted to one or two languages to build translation solutions [26, 38]. The *Sampark* [97] system looks at translation between different pairs of languages in the country as multiple different problems, solved separately, with some commonality of the solution across the pairs in the idea. This pair-wise view of the problem space enables greater incorporation of language expertise into the solution. In the absence of huge resources to train statistical systems, these provided reasonable translations. *Sata Anuvadak* [54], a compilation of 110 independently trained translation systems which used Statistical Machine Translation (SMT) analyzes a multilingual setup through SMT

based approaches. One may attempt to solve multilingual translations through several hops through intermediate pivots. However, error compounds at each step.

Modern-day NMT approaches are capable of handling the sharing of learning between language pairs for which no data is available, through methods of zero-shot learning [36], transfer learning [102], data-augmentation approaches [86, 24]. We are motivated by the success of these.

### 2.3.4 Evaluations and Benchmarks

Workshop on Asian Translation (WAT 2018) [63] tasks on Indian languages had seen models performing lesser on automated evaluation procedures being preferred more by human evaluation. This may lead to interesting explorations on the need of evaluation protocols and data sets for many Indian language translation tasks. We report Bilingual Evaluation Understudy (BLEU) [70], Rank-based Intuitive Bilingual Evaluation Score (RIBES) [34], Adequacy-fluency metrics (AM-FM) [7] for all our attempts and scores from WAT human evaluations when available.

# Chapter *3*

# NMT for English-Hindi

In this chapter, we summarize efforts in studying NMT in a language-pair which is relatively high-resource compared to the others involving the most spoken language in the country - Hindi. First, we build NMT systems for the application case in a narrow domain as described in 3.1. Further, we demonstrate success in building an NMT system which performs well in the general domain, producing state-of-the-art BLEU scores at the time in a public benchmark hosted by the Workshop on Asian Translation (WAT).

## 3.1 Translating a Narrow Domain

### 3.1.1 Introduction

Neural Machine Translation (NMT) has emerged to be the popular choice for end-to-end translation between languages. NMT systems today work surprisingly agnostic to the properties of language at both ends, efficiently capturing linguistic properties through large volume of data.

Indian languages suffer from insufficient digital data, particularly a sentence aligned parallel corpus which current methods in NMT work on. In the absence of enough representative data, NMT systems perform poorly. Thus instead of NMT, we observe that classical Phrase based Statistical Machine Translation (PBSMT) are currently widely prevalent and applied method for Indian languages. NMT based approaches have tried to mitigate the issue of scarcity of data by supplementing the dataset, while still attempting to train a generic translation system. In this chapter, we suggest that acceptable accuracies can be obtained even in the presence of scarcity of data by limiting the NMT system to more specific domains. This hypothesis is supported by the observation that in a narrow domain, the vocabulary used is limited. We are specifically interested in the use of NMT for applications such as translating cricket commentary to Indian languages.

For many practical applications, it suffices to have good translations in a narrow domain. For example, multilingual support services for an organization needs to be concerned primarily about the domain it operates on.

The field of sports is constituted by several such narrow domains that are popular among a large section of people. Many sports commentaries are in English, and recently manual efforts have been initiated to get it into vernacular languages. While English literacy is yet to improve in rural India - these sports and video providers have reached even the remote corners. Cricket is the most popular game in the country, and enough content exists online to attempt the above methods.

In this chapter, we present an effective way to learn from a generic translation system to produce suitable translations for specific domains. Our contributions are:

- Using a generic translation system as a teacher, we propose a simple method to build a superior translation system specific to narrow domains.
- We demonstrate the efficacy of our methods in translating cricket news articles.

The rest of the paper is organized as follows. Section 3.1.2 discuss past work we build on top of and compare against. In section 3.1.3, we motivate the need of the said technique. Section 3.1.4 explains the components of the architecture we employ in detail. We elaborate our experimental setup in section 3.1.5, present our results and discuss them in detail in 3.1.6.

### 3.1.2 Related Work

Generic translation systems work well in learning the syntax of the language and are state of the art at several machine translation tasks. However, when presented with a fresh domain - many out of vocabulary words (OOV) and possibly new syntactic style deteriorate quality. Google's Neural Machine Translation (GNMT) [36] is such a generic system which tries to tackle the problems of low resource languages with scale, learning shared representations by trying to translate multiple languages through the same network.

Attempts of supplementing the training system with language models trained in the domain-specific corpora and using a generic system to generate more synthetic data have also produced promising results [86]. Even with these in place, there are challenges specific to Indian Languages which hinder translation task.

Data scarcity has been an issue for Indian languages MT. Dandapat and Lewis [19] reports the Wikipedia pages for Indian Languages having no more than 125k pages, whereas European languages are close to 1 million pages. Indian Languages fall in the category of morphologically rich languages, wherein inflections and agglutinations lead to exponential growth in vocabulary. Combined, both of the above presents challenges to Indian Languages MT. When a system is trained to translate general corpus on a low resource setting, it has observed to be prone to overfitting - which Dandapat and Lewis [19] remedies by means of early stopping while training. Khan et al. [38] individually addresses language specific features and caveats of running machine translation on Indian languages. Since this work pertains to the *hi-en* pair, we survey the problems pertaining to the same.

Dhariya et al. [22] elaborates on several linguistic differences between English and Hindi which ends up being unfavourable for machine translation systems. When comparing word orderings, English is of Subject-Verb-Object, while Hindi is of Subject-Object-Verb (SOV). Although labelled as SOV, Hindi exhibits free word ordering to some extent, which hampers evaluation metrics based on n-grams. Hindi is morphologically rich and also comprises of several words having largely diverse meanings with changing context.

Recent developments suggest two promising choices for building translation systems for the *hi-en* pair - PBSMT and NMT. PBSMT has extensively been studied for English-Hindi (*en-hi*) pair [23, 51]. Khan et al. [38] demonstrates results SMT based *Moses* on multiple Indian languages to English, which includes the *en-hi* pair.

Several additional NMT based approaches have been tried for *en-hi* pair recently, following the success in many other languages, but on a general domain. Singh et al. [92] experiments and benchmarks two NMT-based approaches - (1) A convolutional sequence to sequence model and (2) a Recurrent Neural Network (RNN) Encoder-Decoder with attention model. Agarwal et al. [1] applies NMT on training data supplementing the model with the use of large monolingual corpora available. Dandapat and Lewis [19], attempting Bangla-English pair uses a bidirectional RNN for an encoder and LSTM for decoder, with enhancements dealing with tradeoff speed and accuracy for production environments.

Domain adaptation techniques have recently been applied to generate speaker personalized translations [59], to yield better results at the granularity of speakers. For *en-hi* pair, domain adaptation has been explored using PBSMT, using news and tourism domains [33]. Koehn and Knowles [43] discusses how an NMT system works very poorly for out of domain data, but this also indicates how NMT works quite well for in-domain data - pointing to the feasibility of building domain specific translation systems.

Pramod Sankar et al. [78] applies alignment to multiple modalities of text based commentary and match videos to segment full-length cricket matches. Our work is towards giving the possiblity of applying translations to the commentary aligned with videos realtime, bringing automation into efforts that are largely manual currently.

Building on top of the above prior work, we motivate the need for specialized translations for narrow domains in the next section.

### 3.1.3 Motivation

**Vocabulary Saturation in Narrow Domain** For a generic translation system, the input and output space complexity increases as the language gets more morphologically rich. But this may not necessarily be the case if we attempt to build focused translation systems for a narrow domain. While sentence structure has to be learned same as the generic translation system - vocabulary and patterns tend to be finite and saturate quickly with lower resource dataset - while maintaning ability to generalize on data from the same domain.

To verify the above intuitions, we conduct a study of the diversity in unique tokens in domains in contrast to how it happens on a typical parallel corpus used to train a generic translation system. For

this, we use the growth of unique n-grams as a proxy to analyze the phenomenon. Our narrow domain here is cricket, while *general* is a news-crawl combined from several domains.

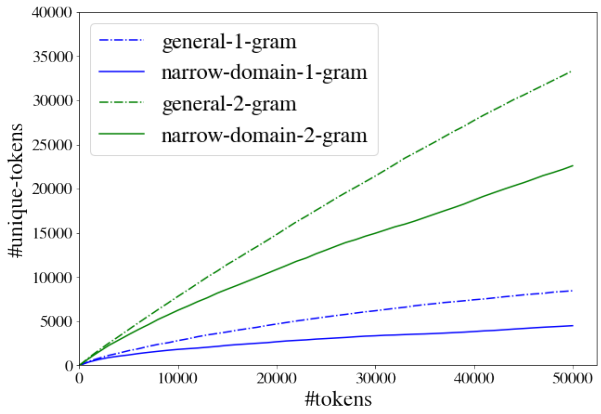


**Figure 3.1** Corpus growth

In figure 3.1, we observe that the complexity of a narrow domain saturates quickly to a value that is lower compared to the general case. This establishes the feasibility of narrow domain translators.

**Ambiguity Reduction in Narrow Domain** Ambiguity in natural language processing is a significant problem. In a general domain, there may be several semantic senses associated with a word. These may not be suited for a narrow domain. Therefore, the need for a focused narrow domain translator does not end at the convenience of lesser input and output classes as argued above. There are nuances in language which enable different meanings of a word in a domain-specific context, compared to a global context of a generic translation approach. Some cases of failure of a generic translator to translate text in a narrow domain are presented in Table 3.1.

| |
|---|
| Sri Lanka enjoyed a productive first morning.<br>श्रीलंका ने पहली बार उत्पादक का आनंद लिया। |
| Pujara got out on a duck.<br>पुजारा एक बतख पर बाहर निकला। |

**Table 3.1** Failure cases of a generic translation system.

In the first example, “morning” is completely missed out. Similarly, the second one brings the impression that a person is riding a duck very distant from a batsman being out on a score zero runs, termed as *duck* in cricket.

**Applications for a Narrow Domain Translation System** The applications and benefits of specific translators for narrow domains are many. Owing to the reduced class space, these tend to be faster than the general-purpose models. Video providers for several sports matches provide content like commentary in multiple languages. Newspapers or magazines today write in their vernacular languages and also in English. Ideally, one would wish for content to be available in all languages.Focused translators can cater to these, enabling a write-once - read-many scenario. We demonstrate such an application through cricket video commentary. A significant figure like a leader of the nation only gives speeches with a fraction of vocabulary than the entirety available - making those narrow domain as well.

Further, we elaborately describe the components we use to build our narrow domain translation system.

### 3.1.4 Neural Machine Translation

A neural machine translation system comprises of a neural network which learns to maximize the probability of the target sequence $\mathbf{y} = \{y_1, y_2, y_3 \cdots y_T\}$ conditioned on the source sequence $\mathbf{x} = \{x_1, x_2, x_3, \cdots x_{T'}\}$. This conditional probability can be expressed as:

$$\log p(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{T} \log p(y_i|y_{<i}, \mathbf{x}) \tag{3.1}$$

#### 3.1.4.1 Encoder - Decoder

An encoder-decoder architecture has been proposed by Sutskever et al. [96] for neural machine translation which reduces the above formulation using two components. An encoder learns to predict a context vector $\mathbf{c} = f(\mathbf{x})$. The encoder $f$ is typically realized by an autoregressive neural network. A decoder learns the probability of the target sequence $\mathbf{y}$ conditioned on $\mathbf{c}$, the context vector provided by the encoder. With the target sequence length denoted using $T$, equation 3.1 can be reformulated as:

$$\log p(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{T} \log p(y_i|y_{<i}, \mathbf{c})$$

One popular choice of architecture for encoder and decoder has been the class of RNNs - usually Long Short Term Memory (LSTM) or Gated Recurrent Units (GRU) for learning longer dependencies. The encoder when realized using an RNN, where the hidden state at a given timestep is denoted by $\overline{h}_t$ and inputs and outputs denoted by $x_t$ and $z_t$, produces a context vector $\mathbf{c}$ as follows:

$$z_t, \overline{h}_t = \text{Encoder-RNN}(x_t, \overline{h}_{t-1})$$
$$\mathbf{c} = \overline{h}_{T'}$$

The decoder modelled by an RNN, now generates the target side representation $\mathbf{y}$ of the source sequence $\mathbf{x}$ as follows:

$$
\begin{aligned}
h_0 &= \mathbf{c} \\
\hat{y}_t, h_t &= \text{Decoder-RNN}(\hat{y}_{t-1}, h_{t-1}) \\
y_t &= \text{softmax}(\hat{y}_t)
\end{aligned}
$$

In this work, we used stacked LSTMs for encoder and decoder.

#### 3.1.4.2 Attention

Attention introduced by Bahdanau et al. [6] enables a decoder in an NMT model to look back into the encoder outputs, to find which ones to give more weightage to while generating a word at a given timestep. Attention can be alternatively viewed as a soft alignment between the current timestep prediction and the encoder outputs. In implementation, this is enabled by using all hidden states of the encoder in deriving the context vector.

$$
\mathbf{c}_t = g(h_1, h_2, \cdots, h_T)
$$

Luong et al. [58] discusses three ways of implementing attention. Our work uses the variant general attention, which accomplishes attending over input as follows. First compute a score for each of the encoder input $\overline{h}_s$ with respect to hidden state at the current timestep $h_t$.

$$
\begin{aligned}
\text{score}(h_t, \overline{h}_s) &= h_t^T \mathbf{W_a} \overline{h}_s \\
\hat{\boldsymbol{a}}_t(s) &= \text{score}(h_t, \overline{h}_s)
\end{aligned}
$$

These can be converted to probability weights of attending to each of the encoder input, by applying softmax. Using the weights, context vector $\mathbf{c}$ is derived as the weighted average of the source input states.

$$
\begin{aligned}
\boldsymbol{a}_t &= \text{softmax}(\hat{\boldsymbol{a}}_t) \\
\mathbf{c}_t &= \sum_s \boldsymbol{a}_t(s) \overline{h}_s
\end{aligned}
$$

### 3.1.5 Experimental Setup

#### 3.1.5.1 Dataset

We start with monolingual data in Hindi comprising of crawled cricket news data, translated through a generic translation system to obtain samples for Hindi-English pair. Sentence aligned pairs obtained constitute the parallel corpus which is our dataset. The dataset is described in Table 3.2.

| Pairs | # English Tokens | # Hindi Tokens |
|---|---|---|
| 123685 | 2299354 | 2410681 |

**Table 3.2** Details of the corpus used.

#### 3.1.5.2 Architecture and Framework

We use OpenNMT-py [40] framework for our experiments. We preprocess and tokenize the input preliminarily using Indic NLP library [1], followed by applying BPE. The vocabulary is brought down by 32000 merge-operations, for source and target while applying BPE. The parallel corpus obtained is split into training, development and testing corpus of fractions 0.6, 0.2 and 0.2 respectively.

We use the following configuration for our model : 2 stacked LSTMs of 500 neurons each of encoder and decoder with general attention from Luong et al. [58] enabled. Embedding layers which maps the input classes to embeddings of size 500 for both the source and target side are used. Beam search decoding is used with a beam width of 5. Once the development set accuracy begins saturating, a learning rate decay of 0.5 onsets, to avoid overfitting. We choose the model which performs best on the development set for testing.

#### 3.1.5.3 Methods

We crawled news articles and used translations from GNMT web service [2] to obtain data to conduct our experiments. The translations from English to Hindi were found to be prone to transliterations at unnecessary word segments of a sentence. Since Hindi already suffers from a scarcity of enough data, inorder to avoid further deterioration from noisy data,the choice of using translations from Hindi to English side was selected, enabling the Hindi text to be more authentic.

The obtained pairs were cleaned to avoid noisy Hindi sources with a large component of English text. Newspaper data contained code-mixed content, primarily tweets embedded within article content. We let the newspaper website navigation instructions stay, as they are valid domain specific data and proper parallel pairs.

#### 3.1.5.4 Evaluations

We take 100 sentences for evaluations and obtain ground truths for the same through 2 human experts.We evaluate and report the standard BLEU metric using the evaluation tools provided by *Moses* [42].

[1] https://github.com/anoopkunchukuttan/indic_nlp_library

[2] http://translate.google.com

### 3.1.6 Results and Discussions

#### 3.1.6.1 Quantitative Results

In Table 3.3, we observe that while GNMT still maintains its advantage in translating to English, our methods have significantly gained in translating from English to Hindi.

| Metric | hi → en | | en → hi | | | |
|---|---|---|---|---|---|---|
| | ours | google | a1-ours | a1-google | a2-ours | a2-google |
| BLEU | 39.25 | **56.28** | **29.38** | 26.07 | **33.71** | 30.06 |
| METEOR | 35.34 | **42.91** | 38.86 | **41.86** | 41.08 | **41.69** |
| ROUGE_L | 64.09 | **71.69** | **56.39** | 55.32 | **60.97** | 57.28 |
| WER | 61.57 | **55.90** | **59.62** | 63.93 | **54.19** | 57.36 |
| CIDEr | 2.89 | **3.77** | **3.04** | 2.82 | **3.31** | 2.99 |
| NIST | 5.38 | **6.24** | **5.51** | 5.45 | **6.18** | 5.82 |

**Table 3.3** Comparison our results with the one provided by generic translation system. BLEU, METEOR, ROUGE_L, WER are reported in percentages.

#### 3.1.6.2 Qualitative Results

In our experiments, we find the free word ordering in Hindi seriously hampering BLEU based evaluations with limited references. In several locations, the model has come close to the semantic notions in the source sentence - while the output style turns out to be different than the teacher's predictions.

Qualitative results for both Hindi to English and English to Hindi sides are presented in figures 3.4 and 3.5 respectively. We demonstrate where our models have gained success compared to the teacher and where they have failed.

**English → Hindi** We observe the context of cricket helping out in translating correctly in several samples. In the first sample, we see the general model confusing "test" with an examination instead of the test matches from cricket. Similarly, in the second example - we find that the model predicts getting out as "getting out of a place", missing out that it is "getting out" from an innings in the cricket context. Series is mistaken to be the literal series, not the named entity cricket series - providing an unnatural feel to the translation. In the final sample, we see that the general model has failed to capture the content of the sentence - while our fine-tuned model has come very close.

| tag | success |
|---|---|
| source | Yuvraj and Raina scored less than 19.5 in this test. |
| g-pred | इस परीक्षा में युवराज और रैना ने 1 9 .5 से भी कम स्कोर बनाए। |
| pred | युवराज और रैना ने इस टेस्ट में 19 से 5 से कम रन बनाए । |
| source | Both openers Jason Roy (00) and Alex Hales (14) got out quickly. |
| g-pred | दोनों सलामी बल्लेबाज जेसन रॉय (00) और एलेक्स हेल्स (14) जल्दी बाहर निकले। |
| pred | सलामी बल्लेबाज जेसन रॉय ( 00 ) और एलेक्स हेल्स ( 14 ) जल्दी आउट हो गए । |
| source | He said that the series should continue. |
| g-pred | उन्होंने कहा कि श्रृंखला जारी रखना चाहिए। |
| pred | उन्होंने कहा कि सीरीज आगे बढ़ना चाहिए । |
| source | Kerala's Sanju Samson came to the headlines playing at Rajasthan Royals in IPL. |
| g-pred | केरल के संजू सैमसन आईपीएल में राजस्थान रॉयल्स में खेलने वाली शीर्षकों पर आए। |
| pred | आइपीएल में राजस्थान रॉयल्स की तरफ से खेलने आए केरल के संजू सैमसन सुर्खियों में आए । |

| tag | failure |
|---|---|
| source | Rohit would like to keep records |
| g-pred | रोहित रिकॉर्ड रखना चाहते हैं |
| pred | रोहित को करना चाहेंगे |
| source | The player runs between two lines in a row and when the beep rings it must turn . |
| g-pred | खिलाड़ी लगातार दो लाइनों के बीच दौड़ता है और जब बीप बजती है तो उसे मुड़ना होता है । |
| pred | एक लिहाज से इस खिलाड़ी के बीच रनों का जो खिलाड़ी आया वो कब और कब टर्न करनी चाहिए । |
| source | Only a few captains get to that place . |
| g-pred | उस स्थान पर बहुत ही कम कप्तान पहुंच पाते हैं । |
| pred | सिर्फ एक कप्तान को यह जगह मिलता है । |
| source | Economy of the best |
| g-pred | सबसे बेस्ट रही बुमराह की |
| pred | बेहतरीन बल्लेबाजी का |

**Table 3.4** Qualitative Results and Comparisons: English to Hindi. g-pred is the prediction of a generic system, while pred is our prediction.

| tag | success |
|---|---|
| source | बच्चे युवराज के साथ समय बिताकर काफी खुश थे । |
| g-pred | The children were very happy to spend time with the Prince . |
| pred | The children were very happy with respect with Yuvraj Singh . |
| source | अब विश्वकप के लिये बस 16 महीनों का ही समय बचा है । |
| g-pred | Now the time for the World Cup is only 16 months left . |
| pred | Now only 16 months has left for the World Cup . |
| source | इस बार फाइनल में मुंबई का सामना पुणे से था । |
| g-pred | This time Mumbai was in the finals from Pune . |
| pred | This time Mumbai faced Pune in the final . |
| source | उन्होंने बोल्ट को पुजारा के हाथों कैच आउट कराया । |
| g-pred | He catches the ball to Pujara ’ s hand . |
| pred | He got Bolt caught by Pujara . |

| tag | failure |
|---|---|
| source | आपको बता दें कि युवराज सिंह को 2011 वर्ल्ड कप के बाद कैंसर होने का पता चला था । |
| g-pred | Tell you that Yuvraj Singh was diagnosed with cancer after the 2011 World Cup . |
| pred | Let me tell you that Yuvraj Singh was found to be married after the 2011 World Cup . |
| source | नवरोज 42 के कुल योग पर आउट हुए । |
| g-pred | Nowrose was out on 42 . |
| pred | The total of 42 was dismissed . |
| source | बुमराह के बाद सबसे अच्छी इकोनॉमी स्पिन गेंदबाज़ यजुवेंद्र चहल की रही । |
| g-pred | After Bumrah , the best economy spin bowler was Yajuwendra Chahal . |
| pred | The best economy bowlers after Bumrah was the master of spin bowlers . |
| source | वहीं घरेलू सरजमीं पर उसे 264 टेस्ट में 98 में जीत और 52 में हार मिली है । |
| g-pred | On the domestic front , he won 98 of the 264 Tests and won 52 . |
| pred | At the same time , he has lost in a lot of Tests and lost in 52 Tests . |

**Table 3.5** Qualitative Results and Comparisons

**Hindi → English** Consider the first example of success in 3.5, where the generic model has taken "Yuvraj" to be "Prince", while our model being domain-specific figured it is required to map to "Yuvraj Singh", the player. In the third example, our model is able to distinguish "Pune" as the team, unlike the generic model which takes it to be a place. In the final example, we see cricket specific language coming in at the target, due to the influence of the domain-specific training. The other prediction turns out to be close enough, provided one takes cricket out of the context.

Among the failure tagged samples, we observe that the model struggles when the vocabulary is outside of what it has learnt. In the first sample, it confuses "marriage" with "cancer", while getting the remaining parts - which are relevant to cricket domain correct. In the second sample, it misses out "Nowrose" - not having seen it in the training corpus, which applies again "Yajwendra Chahal" in the third sample.

### 3.1.7 Cricket Commentary


**Figure 3.2** Scenes from a cricket match overlayed with subtitles. Possible application of this work. English overlays are shown in the first row, and the corresponding translated Hindi commentary below.

We now further explain the application in the specific case of cricket commentary. Much of the cricket content is not available in regional Indian languages such as Malayalam or Punjabi. However, there is significant following for the sport in these places. Making available the cricket commentary in regional languages will enable accessibility for a widely enjoyed sport.

Towards this end, we are using the proposed narrow domain translation system. The method works by first using a service to access cricket commentary information from widely followed websites (such as ESPNcricinfo.com). These provide live commentary feed for content. One implementation of this system can be obtained by translating this content using the proposed narrow domain NMT system in real-time. This content can be then overlaid over the ongoing television feed.

This kind of system would enable detailed match analysis and expert commentary to be made available to an Indian viewer in their corresponding regional language.

We illustrate results from a preliminary prototype for such a system in the figure 3.2. We have specifically considered adding a black overlay area in which we render the text. As can be seen from the figure, we show examples of the English and corresponding Hindi text being rendered for the specific frame. As most of the rendering and machine translation as well as fetching of information from website can be performed in real-time or near real-timE, WE believe that such a system can therefore be feasible. Further, the quality of translation obtained from the proposed narrow domain NMT system is observed to be satisfactory to enhance the viewer experience for this use case.

### 3.1.8 Conclusions and Future Work

In this work, we propose a method for narrow domain neural machine translation. Our method is trained for a narrow domain corpus using data obtained from a generic NMT system. This improves performance in the narrow domain over the source generic NMT system. This method can be further improved by using some amount of ground truth translation corpus, which is feasible for the narrow domain. We also provide a set of human-annotated examples that are used for evaluating quality for the narrow domain.

Our method is motivated by the use case of translating cricket commentary data. Thus the proposed system is specifically trained on a corpus of cricket data. For this application, we show that the proposed system performs satisfactorily through qualitative and quantitative evaluation. This system would also be practically useful as cricket is a widely followed sport, and through the proposed system one can automatically obtain narrow domain NMT systems for the various regional Indian languages.

## 3.2 Translating General Domain

In this section, we describe expanding the datasets from the previous cricket-based experiments and make use of massive data augmentation approaches in modern machine learning, to build a general-purpose machine translation system for English-Hindi language pair. We describe ahead the details associated with the tokenization, architecture and data augmentation methods used. These are the three components that helped in obtaining superior results on the corpus provided by the organizers of WAT-2018.

### 3.2.1 Components

#### 3.2.1.1 Tokenization

A popular method of addressing rare-words without compromising coverage of the entire corpus was Byte Pair Encoding (BPE) [87], which used a deterministic greedy compression based algorithm to bring the vocabulary down to a finite feasible value.

SentencePiece [46] builds on top of byte pair encoding. Unlike BPE, which is agnostic to language, SentencePiece gives the most likely derivation of a sentence composed of subword units. This setting reduces to character level in case a completely unknown sentence/word is provided, and the translation model also learns to transliterate. We use SentencePiece for its merits mentioned above.

#### 3.2.1.2 Convolutional Sequence to Sequence Learning

In our submission, we employ the Convolutional Sequence to Sequence architecture (ConvS2S) [27]. ConvS2S follows an encoder-decoder architecture. This has the advantage of being faster than the popular Recurrent Neural Network (RNN) based encoder decoder architectures with attention. This is because the context is built through multiple inputs stacking $k$ convolution blocks ($O(\frac{n}{k})$) with the ability to build in parallel representations for multiple parts of the sentence, unlike through time in the RNN ($O(n)$).

A 1-D convolutional filter of width $w$ with two channels at the output sliding over the embeddings of the text inputs constitute a basic convolutional block. Output of one channel builds up context representation and the other is used to enable gating through Gated Linear Units (GLUs) [20]. The encoder is constructed by stacking $k$ of the above setup, creating a receptive field controlled by $w$ and $k$. The decoder is similar to the encoder in architecture, with a fully connected layer projecting output to vocabulary size.

In the next section, we describe how the components explained above are implemented and used in training - including generating dataset, preprocessing and filtering the training samples, hyperparameters of the architectures in place and evaluations.

### 3.2.2 Experimental Setup

#### 3.2.2.1 Dataset

In our experiments, we use the training data provided by organizers. In addition, we also use data obtained from translated Hindi content available on Internet. Top level statistics of the data used are provided in Table 3.6.

| **Dataset** | **Pairs** | **Tokens** | |
|---|---|---|---|
| | | hi | en |
| IITB train | 1,492,827 | 22.2M | 20.6M |
| IITB train$^{\dagger}$ | 923,377 | 20.3M | 18.9M |
| National News | 2,495,129 | 41.2M | 39.0M |
| Backtranslated | 5,653,644 | 77.5M | 91.9M |
| IITB dev | 505 | 10,656 | 10,174 |
| IITB test | 2,507 | 49,394 | 57,037 |

**Table 3.6** Descriptions of the corpora used, IITB train$^{\dagger}$ is a filtered version of the IITB train corpus.

The training corpus provided by the organizers, hereafter denoted by IITB-corpus consists of data from mixed domains. There are roughly 1.5M samples in training data from diverse sources, while the development and test sets are from newspaper crawls. In addition to this, monolingual data collected by the organizers from several sources are used in our backtranslation enabled attempts at training an NMT system. There are 45M samples in the monolingual corpus provided.

We enhanced the training data with additional pairs, but automatically translated. Note that no manual translation was used to create additional data. We obtain 2.5M Hindi sentences automatically translated to English from newspapers and similar resources, obtained from Internet. This data is some what domain specific. They are primarily, from news articles related to national news. This is mentioned as National News in Table 3.6.

We also create a parallel corpus through backtranslation using the organizers monolingual Hindi data hereafter denoted by Backtranslated, the details of which are also included in Table 3.6 and the methods of creation elaborated in Section 3.2.2.3.

#### 3.2.2.2 Data Processing

We train separate SentencePiece models using official implementation available online [3] with vocabulary restricted to 8000 units to function as a learned tokenizer for both English and Hindi. We use the unigram model, which gives language aware tokenization.

[3] https://github.com/google/sentencepiece

| Dataset | en-hi | | | | hi-en | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU | RIBES | AM-FM | Human | BLEU | RIBES | AM-FM | Human |
| IITB train† | 13.25 | 0.695113 | 0.647220 | - | 11.83 | 0.675462 | 0.572900 | - |
| National News | 18.77 | 0.748008 | 0.697630 | - | 19.53 | 0.745764 | 0.614260 | - |
| +IITB train† | 19.69 | 0.758365 | 0.699810 | 69.50 | 20.63 | 0.751883 | 0.623240 | 72.25 |
| Backtranslated | 16.77 | 0.714197 | 0.664330 | 50.50 | - | - | - | - |
| 2017 Best | 21.39 | 0.749660 | 0.688770 | 64.50 | 22.44 | 0.750921 | 0.629530 | 68.25 |
| 2018 Best | 20.28 | 0.761582 | 0.704220 | 77.00 | 17.80 | 0.731727 | 0.611090 | 67.25 |

**Table 3.7** Quantitative results of translating English to Hindi and vice versa.

To filter any noisy content from IITB corpus, *langdetect*[4] and removed every pair which had probability of being in the respective language less than 0.95. This gave us roughly 0.92M pairs for training, from IITB corpus and is indicated as IITB train† in Table 3.6. English data is kept true-cased, which we found to have better results consistently with our NMT model.

#### 3.2.2.3 Training

In our experiments we use the fairseq [5] toolkit. For the tasks in this submission we use the ConvS2S model. The encoder and decoder embeddings have a dimension of 512. The hidden units in the encoder and decoder are also 512 dimensional, following [27]. We use convolutional filters of width 3 and 20 layers stacked for both the encoder and decoder. A dropout with probability 0.1 is put in-place right after the embeddings layer for better generalization. The training is run in batches of maximum 4000 tokens at a time, which is on an average 140 sample sentences per batch. The model is trained to minimize the categorical cross-entropy loss at the token level using Nestorov accelerated gradient descent. Decoding is performed through beam search with a beam width of 10.

We run training using four NVIDIA 1080Ti-s until validation loss hasn't improved for 3 epochs straight. The training time was roughly 2 days and stopping around 30-40 epochs. We keep our model hyperparameters constant as specified across experiments and work with different combinations of corpora created from augmenting the National News dataset and official parallel corpora. For creating the Backtranslated corpus, we use a model trained to translate from Hindi to English using both National News and IITB corpus. We filter the obtained pairs using confidence of translation obtained from the beam-score and further to pairs with a length between 10 and 30 tokens.

[4] https://github.com/Mimino666/langdetect

[5] http://github.com/pytorch/fairseq (formerly fairseq-py)

#### 3.2.2.4 Evaluations

We report Bilingual Evaluation Understudy (BLEU) [70], Rank-based Intuitive Bilingual Evaluation Score (RIBES) [34], Adequacy-fluency metrics (AM-FM) [7] for all our attempts and scores from WAT-2018 human evaluations(Human in Table 3.7) when available.

BLEU is computed as the geometric mean of unigram, bigram, trigram and 4-gram precision multiplied by a brevity penalty (BP). BLEU ranges from 0 to 1, but the values reported in Tables 3.7 and 3.8 are in percentages. RIBES, also giving a value in $[0, 1]$ was proposed to tackle shortcomings of BLEU in distant language pairs, where changes in word ordering deteriorates BLEU.

#### 3.2.2.5 Discussions

The results using our systems for WAT-2018 are presented in Table 3.7 (see some additional results in Table 3.8). The first part of the table consists of results on combinations of datasets and augmentations. All values are for models trained from scratch. In the second part, the current leader board is indicated for comparison. Note that entries in this part don't correspond to a single submission, but the values corresponding to the best in the respective metric.

Our submission based on the combination National News and IITB corpus tops human evaluation in Hindi to English, and ranks second in English to Hindi. We demonstrate the possibility of distilling knowledge of online available sources into a usable translation model. We successfully use the ConvS2S architecture along with SentencePiece to obtain results comparable to the top submissions. Our experiments also indicates data augmentation using backtranslation positively works for the Hindi-English pair.

### 3.2.3 Transformer Experiments

In this section, we present a set of experiments and results post WAT-2018 involving the Transformer Architecture [98]. Two variants of the architecture - Transformer Base and Transformer Big outperformed then state of the art ConvS2S models in the WMT German-English and French-English translation tasks.

We used the Transformer-Base architecture in further experiments with the National News + IITB corpus where ConvS2S performed the best, with the rest of the pipeline being kept same as described before. We went with the default hyperparameters provided by *fairseq* framework - which did not give us impressive results.

Following Popel and Bojar [75], we modified the hyperparameters for initial warm-up steps of 16000 without any learning rate decay, starting from a learning rate of 0.25, followed by an exponential decay of learning rate. We also had to enable delayed gradient updates [68] to simulate a larger batch on smaller GPU before the model demonstrated any learning. During inference time, we averaged checkpoints of the model at different epochs once the loss on the development set had plateaued to obtain better results than a single checkpoint.

| Architecture | BLEU | RIBES | AM-FM |
|---|---|---|---|
| ConvS2S | 19.69 | 0.758365 | 0.699810 |
| Transformer | 21.10 | 0.771549 | 0.712200 |
| +Averaging | **21.57** | **0.773923** | **0.712110** |

**Table 3.8** Transformer-Base vs ConvS2S on National News + IITB corpus, for English to Hindi direction.

In Table 3.8, we compare the performance of the transformer with that ConvS2S. Consistent with observations in languages like German-English and French-English, the transformer network produces better results than ConvS2S on all metrics. The averaged model performs the best in all metrics in English to Hindi translation task, at the time of writing this paper.

## 3.3 Conclusion

In this chapter, we demonstrated the success of NMT systems in both general-domain and a narrow-domain for the English-Hindi language-pair. Further enhancements in Neural MT could provide high-performing translation functionality in other Indian languages, many of which are low-resource in terms of available digital content. In the next chapter, we study extending the advancements to more languages spoken in the subcontinent.

# Chapter *4*

# Multilingual Approaches for Low Resource languages

In this chapter, we extend the previous model incorporating multilingual formulations following [36] and demonstrate the efficacy in many low-resource Indian languages. We introduce the Mann Ki Baat corpora, aligned between 7 languages and demonstrate success of the multilingual model in the corpora.

## 4.1 Datasets

In this work, we compile data available from several sources, each of which we describe below. The languages we use in our experiments include English (en), Hindi (hi), Bengali (bn), Malayalam (ml), Tamil (ta), Telugu (te) and Urdu (ur). Summary of the corpora used is provided in Table 4.1.

| | bn | hi | ml | ta | te | ur |
|---|---|---|---|---|---|---|
| wat-train | 337K | 84K | 359K | 26K | 22K | 26K |
| iitb-train | - | 1.5M | - | - | - | - |
| ilci-train | 50K | 50K | 50K | 50K | 50K | 50K |
| wat-dev | 0.5K | 0.5K | 0.5K | 0.5K | 0.5K | 0.5K |
| iitb-dev | - | 0.5K | - | - | - | - |
| iitb-test | - | 2.5K | - | - | - | - |
| wat-test | 1K | 1K | 1K | 1K | 1K | 1K |
| wat-mono | 453K | 105K | 402K | 30K | 24K | 29K |
| wiki-mono | 371K | - | 1M | 1.6M | 2.66M | - |
| iitb-mono | - | 19.2M | - | - | - | - |

**Table 4.1** WAT Corpus Details

**Parallel Data** We primarily rely on the IITB-hi-en as parallel data for our training. We also use an extended version of our system from WAT-2018 [72] to backtranslate the available monolingual data in Hindi to English, and to augment the training. To take full advantage in the multilingual setting, we use the ILCI corpus to include training pairs in all directions. In addition to the above, we use WAT-ILMPC, which gives ability to the multilingual model to implicitly pivot through English. The designated splits for testing and validation (development) for IITB-hi-en and WAT-ILMPC are set aside for evaluations, while the remaining pool of data is used in training.

***Mann Ki Baat* test set** We extract a new multilingual test set for Indian languages, using data available online[1]. These are transcriptions of Prime Minister's address to the nation manually translated into multiple languages of the country. Our objective in creating this test-set is to find how well the models evaluated on WAT-ILMPC is able to generalize to a new data/situation.

**Monolingual Data** We use Wikipedia text to create samples to enhance model through backtranslation. The IITB-hi-en and WAT-ILMPC come with their share of monolingual data as well, which we make good use of.

## 4.2 Multilingual Neural Machine Translation

In a multilingual setting, we have several languages to translate to and from. Hereafter we use *xx*, *yy* and *zz* as placeholders to denote arbitrary languages. The languages and the respective notations (shown in the bracket) as English (en), Hindi (hn), Telugu (te), Tamil (ta), Malayalam (ml), Urdu (ur), Bangla (bn).

Our multilingual setup closely follows Johnson et al. [36]. We share the encoder and decoder parameters across all languages, following the fact that the vocabulary is common on both sides and the representations being the same relaxes the computation. A control token is prepended to the input sequence to indicate which direction to translate to (`__t2xx__`). The decoder learns to generate the target given this input.

One advantage of using multilingual setups is that it enables zero shot learning. Few shot learning (such as zero-shot and one-shot) has emerged as an effective tool for addressing the lack of training data in many practical problems in computer vision [94, 101], language processing [36] and speech processing [95]. Pairs in language directions that did not exist in the training set can still be translated. They can also aid low resource settings of Indian languages. Rapid adaptations to new languages can also be done through multilingual setups [65]. Aharoni et al. [2] perform experiments at an even larger scale to find that massive-multilingual setups leading to large improvements in low resource setting.

[1] https://www.narendramodi.in/mann-ki-baat

Thanks to *Bahubhashak* initiative for bringing the attention to this, as a language resource.

| | |
|---|---|
| source | The Führer isn't recruiting you as soldiers, he's looking for a secretary. |
| target | यह आवश्यक नहीं है. फ्यूरर तुम को सैनिकों के रूप में भर्ती नहीं कर रहा है, वह एक सचिव खोज रहा है. |
| mm | फ्यूरर आप सैनिकों के रूप में भर्ती नहीं कर रहा है, वह एक सचिव की तलाश कर रहा है. |
| mm+ft | फ्यूरर आप सैनिकों के रूप में भर्ती नहीं कर रहा है, वह एक सचिव की तलाश में है। |
| source | No man in the sky intervened when I was a boy to deliver me from Daddy's fist and abominations. |
| target | आकाश में कोई भी आदमी हस्तक्षेप किया जब मैं एक लड़का था मेरे पिताजी की मुट्ठी और घिनौने काम से वितरित करने के लिए। |
| mm | आसमान में किसी आदमी ने तब हस्तक्षेप नहीं किया जब मैं पिताजी की मुट्ठी और घिनौनी से मुझे पहुंचाने वाला लड़का था। |
| mm+ft | आकाश में कोई भी आदमी हस्तक्षेप किया जब मैं एक लड़का था पिताजी की मुट्ठी और अपमान से मुझे देने के लिए. |
| source | Tonight, we offer you the one illusion that we dreamed about as kids but never dared perform until now. |
| target | आज रात, हम तुम हम बच्चों के रूप में के बारे में सपना देखा था कि एक भ्रम पेशकश... ... लेकिन अब तक प्रदर्शन करने की हिम्मत कभी नहीं. |
| mm | आज रात हम आपको एक भ्रम की पेशकश करते हैं कि हम बच्चों के रूप में सपने देखा लेकिन अब तक कभी हिम्मत नहीं की प्रदर्शन। |
| mm+ft | आज रात, हम तुम्हें एक भ्रम की पेशकश हम बच्चों के रूप में के बारे में सपना देखा कि लेकिन अब तक प्रदर्शन करने की हिम्मत कभी नहीं की. |
| source | What if I tell you it's got a Foot Locker on one side and a Claire's Accessories on the other. |
| target | क्या होगा अगर मैं तुम्हें बता मिल गया है एक फुट एक तरफ लॉकर और दूसरे पर एक क्लेयर सहायक उपकरण. |
| mm | क्या अगर मैं तुम्हें बताता हू कि यह एक तरफ एक फुट लॉकर और एक क्लेयर एक्सेसरीज मिल गया है. |
| mm+ft | क्या मैं यह एक तरफ एक फुट लॉकर मिल गया है और एक क्लेयर के एक्सेसरीज पर. |
| source | I could take up Jon's duties while he's gone, My Lord. |
| target | जबकि वह, चला गया मेरे प्रभु मैं जॉन के कर्तव्यों का समय लग सकता है। |
| mm | मैं जॉन के कर्तव्यों को संभाल सकता था जब वह गया है, मेरे प्रभु. |
| mm+ft | मैं जॉन के कर्तव्यों को संभाल सकता था जब वह चला गया है, मेरे प्रभु। |

**Table 4.2** Five example sentences in English and their translations in Hindi. Readers may note the (i) length (ii) complexity and (iii) diversity among models of the sentences

### 4.2.1 Evaluation Protocol

We report Bilingual Evaluation Understudy (BLEU) [70] on all the test sets discussed previously. BLEU is a precision based metric for automatic evaluation of machine translation test sets.

### 4.2.2 Results

In this section, we present our results on WAT-ILMPC, IITB-hi-en and our newly explored *Mann Ki Baat* test-sets. In the process, we attempt to comprehensively study the performance.

#### 4.2.2.1 On IITB-en-hi and WAT-ILMPC

We have now one-model yielding all-round performance on two tasks — one domain specific and one generic test set being trained on a combination of both datasets. We report the results on the automated evaluation procedures for the same here. A summary of our numbers on the BLEU evaluation metric for each task is presented in Table 4.3, across two variants of the same model we have.

| direction | model | wat-bn | wat-hi | wat-ml | wat-ta | wat-te | wat-ur | iitb-hi |
|---|---|---|---|---|---|---|---|---|
| from-en | *ILMULTI* | 10.21 | 20.72 | 8.17 | 12.55 | 19.10 | 16.56 | 20.17 |
| | + fine-tune on WAT | 12.26 | 26.25 | 9.22 | 16.06 | 23.63 | 19.51 | 18.31 |
| to-en | *ILMULTI* | 17.35 | 29.74 | 12.45 | 17.63 | 25.56 | 21.03 | **22.62** |
| | + fine-tune on WAT | 19.58 | 31.55 | 14.77 | 21.94 | 30.91 | 24.60 | 20.66 |

**Table 4.3** BLEU scores on WAT-ILMPC (wat-*xx*) and IITB-hi-en (iitb-*xx*) test sets. Directions are indicated in the first column. Our model is the best in English-to-Hindi, in the top 3 in Hindi to English in IITB-hi-en. In addition to this, it performs well on WAT-ILMPC evaluation metrics.

Our base model i.e., *IL-MULTI* obtains the best results till date for the English to Hindi direction task for IITB-en-hi. The results for Hindi to English is also quite comparable.

For WAT-ILMPC tasks, our system gives comparable numbers while translating to English from other available languages. However, in the other direction, we observe that the results are lower. Attempts to fine-tune and adapt *IL-MULTI* to the domain specific subtitles dataset gives us *IL-MULTI+ft* , which demonstrates a boost in the evaluation scores on test sets specific to WAT-ILMPC. However, we observe performance degrades in the IITB-hi-en test set, which is general domain.

The overall ordering in the evaluation metrics show some patterns and consistency. While *xx-en* is giving comparable performances in many languages to the best values, *en-xx* fails to show similar performance. This could be due to imbalances in training samples among language-directions. The WAT-ILMPC dataset heavily pivots through English, the training data being strictly with English on

one side and other languages on the other. This leaves training with a situation where there is a lot of data for the decoder to learn to generate English targets, while proportionately very less in languages of smaller resources. We try to solve this scenario by augmenting training with heavy monolingual data on the decoder, trying to translate from noisy backtranslated sources in English or Hindi. The model thus warm-started from *IL-MULTI* , *IL-MULTI+bt* gives us increase in BLEU for the less resource directions.

Our results on WAT-ILMPC and IITB-hi-en puts us in a comparable setting for many language pairs. However, our multilingual training enables us to study how many non *xx-yy* directions work, where *xx* or *yy* need not necessarily be English. For this, we use the *Mann Ki Baat* dataset to obtain a proper multilingual dataset, using which we report our numbers in these directions below.

#### 4.2.2.2 Multilingual Results

For our multilingual results, we present BLEU scores after using our *SentencePiece* models for tokenization. Though non-standard, this compensates for agglutinative languages with free word orderings ending up with low BLEU scores albeit translations being of reasonably good quality. Since the tokenization setup is uniform across languages, this lets us compare the several directions possible fairly.

| srcs | bn | en | hi | ml | ta | te | ur |
|---|---|---|---|---|---|---|---|
| bn | 99.87 | 15.20 | 14.43 | 8.15 | 3.85 | 7.35 | 0.0 |
| en | 9.50 | 100.00 | 15.37 | 8.71 | 4.60 | 8.20 | 0.0 |
| hi | 11.97 | 21.79 | 99.20 | 9.09 | 5.83 | 9.49 | 0.0 |
| ml | 6.85 | 12.00 | 9.84 | 99.90 | 3.37 | 7.32 | 0.0 |
| ta | 3.51 | 6.92 | 5.86 | 4.06 | 99.91 | 3.63 | 0.0 |
| te | 6.99 | 11.06 | 9.54 | 8.00 | 3.59 | 99.74 | 0.0 |
| ur | 0.00 | 0.00 | 19.55 | 0.00 | 0.00 | 0.00 | 100.0 |

**Table 4.4** BLEU scores over translation pairs in Mann Ki Baat multilingual test-set. The complexity of sentences are high, with long and rich sentences of the kind included in ministerial speeches.

In Table 4.4, we present results from *Mann Ki Baat*, using our model *IL-MULTI+bt* , which resulted the best numbers overall. The results indicate improvements especially in generating translations the resource scarce languages for a rich source. Our English-Hindi models give strongest performance, as in the setting above. The rows indicate source language and the columns target language.

From the grid, we see performance in (near) zero shot learning for language pairs which were not represented well enough in the dataset. The results in *ml-te* and *bn-te* show comparable performance with *hi-te* and *en-te*. The same phenomenon exists while translating to Tamil and Malayalam. Almost

all the models have learned to copy well, i.e, when the source and target languages are same. It gets the target accurate $\sim$ 100% of the time.

In our experiments we find that *IL-MULTI+ft* which performed in the WAT-ILMPC test sets better failed to show good results here, indicating that the model is clearly adapting to WAT-ILMPC data.

### 4.2.3 Qualitative Examples and Discussions

In Tables 4.5 and 4.6, we illustrate the success and failure cases among a few languages. Our model has managed to capture meaning most of the time. When the ground-truth is incorrect due to the subtitle corpus being noisy, our model yet manages to give the correct translation (WG -3 for example). There are samples like music tracks which are not translated and stay the same. While these are acceptable in a subtitle setting it is not so for pure translation attempts.

Sometimes the ground truth, being subtitles seem to have sentences for which word-meanings can not be inferred from the given context. Take WB-4 in Table 4.6 for example. The ground truth has a lot of context, assuming the person has a disease and is recovering while translating. It can be noticed that our models confuse it to be improvement or betterment, with the context being absent in the source.

Our model is not able to keep up to the mark while translating poetic sentences. It seems to be missing out on colloquial sentences as well, translating almost literally, missing intent although capturing enough meaning, as in WB-5.

## 4.3 Discussions

It is important to isolate errors or noise and improve the evaluation standards over time to enable more authentic evaluations. In the previous section, we illustrated several qualitative examples where our model was working fine, but the ground truth itself is incorrect, incomplete or has extra information. In this section, we take one iteration at improving the state of things of the test set. down a few errors and discuss our take on the evaluation metrics.

**Shortcomings in the Evaluation Set** In our experiments, we found our multilingual models trained on large data to be robust and well performing for translations, but the BLEU scores obtained during evaluation indicates otherwise. Our best results are obtained after fine-tuning a general-purpose model on the training set of the WAT-ILMPC. Further inspections indicate the ground truth samples to be incomplete, at either sides or both. In Table 4.7, we illustrate a few of these.

### 4.3.1 More on WAT-ILMPC

In the main paper, the content was focused on building a general-purpose machine translation system among 7 languages. We argued in favour of keeping *IL-MULTI* for future experiments against *IL-*

| | | |
|---|---|---|
| WG -1 | wat.ml | അവർ ഒരു രോഗകാരിയെ പടർത്തുന്നതായി ഞാൻ കരുതുന്നു, അവർക്ക് ആരോഗ്യകരമായ ഒരു ഹോസ്റ്റ് വേണം. |
| | wat.en | I think they're spreading a pathogen, and they need a healthy host. |
| | mm.en | I think they're spreading a sick person, they need a healthy host. |
| | mm+ft.en | I think they're spreading a patient, they need a healthy host. |
| WG -2 | wat.en | One, two, three, four, five, six, seven... eight, nine, ten, eleven. |
| | wat.ml | ഒന്ന്, രണ്ട്, മൂന്ന്, നാല്, അഞ്ച്, ആറു, ഏഴ്... എട്ട്, ഒമ്പത്, പത്ത്, പതിനൊന്ന്. |
| | mm.ml | ഒന്ന്, രണ്ട്, മൂന്ന്, നാല്, അഞ്ച്, ആറ്, ഏഴ്... 8, 9, 10, പതിനൊന്ന്. |
| | mm+ft.ml | ഒന്ന്, രണ്ട്, മൂന്ന്, നാല്, അഞ്ച്, ആറ്, ഏഴ്... എട്ട്, ഒമ്പത്, പത്ത്, പതിനൊന്ന്. |
| WG -3 | wat.en | Come on listen to this corporal for a moment sir if he's not dead already |
| | wat.ml | Come on. സർജെന്റ്റ് , ഇയാൾ പറയുന്നത് ഒന്ന് കേൾക്കു. സർജെന്റ്റ്, ഇവിടെ മാരകമായി പരുക്കേറ്റ ഒരു പട്ടാളക്കാരനെയും കൊണ്ട് വന്നതാ. |
| | mm.ml | ഒരു നിമിഷം ഈ കോർപ്പറൽ പറയുന്നത് കേൾക്കു സർ, അവൻ മരിച്ചിട്ടില്ലെങ്കിൽ |
| | mm+ft.ml | ഈ കോർപ്പറൽ ഒരു നിമിഷം കേൾക്കു സർ... ...അവൻ മരിച്ചിട്ടില്ലെങ്കിൽ |
| WG -4 | wat.hi | तुम मेरे घर क्यों आये थे? |
| | wat.en | Then why do I hurt so bad right now? |
| | mm.en | Why did you come to my house |
| | mm+ft.en | Why did you come to my house |
| WG -5 | wat.hi | मुझे लगता है कि यह वास्तव में किसी तरह की एक उड़ान बात थी. |
| | wat.en | I think it was actually a flying thing of some kind. |
| | mm.en | I think it was actually a kind of flying thing. |
| | mm+ft.en | I think it was actually a kind of flying thing. |
| WG -6 | wat.hi | 16 दिसम्बर को हमारा काम कुछ-कुछ घुड़सवार सेना जैसा होगा, ओके? |
| | wat.en | Now, our role on the 16th of December is to be a bit like the cavalry, OK? |
| | mm.en | On December 16, our work will be like some horsemen army, OK |
| | mm+ft.en | On December 16th, our job will be like some riding army, okay |
| WG -7 | wat.en | I don't know how you find the time to raise kids and teach. |
| | wat.ta | குழைந்தைக்கு கத்துகொடுக்குரதுக்கு எப்டி உங்களுக்கு நேரம் கெடச்சுது? |
| | mm.ta | நீங்கள் குழந்தைகள் வளர்க்கும் மற்றும் படிக்கும் நேரம் எப்படி தெரியாது. |
| | mm+ft.ta | நீங்கள் குழந்தைகள் வளர்க்கும் மற்றும் படிக்க நேரம் கண்டுபிடிக்க எப்படி எனக்கு தெரியாது. |

**Table 4.5** Success Cases from WAT test-set.

| | | |
|---|---|---|
| WB-1 | wat.ml | അത് ഇനി നിൻറെ ഹൃദയത്തിൽ എത്ര വെറുപ്പും ദേഷ്യവും ഉണ്ടെന്ന് പറഞ്ഞാലും ശരി.. കാഞ്ചി വലിക്കാനുള്ള സമയമാകുമ്പോൾ.. മിക്ക ആൾക്കാർക്കും അതിന് കഴിയില്ല.. |
| | wat.en | No matter how much hate and anger you may have in your heart, when it comes time to pull the trigger most people can't do it. |
| | google.en | No matter how much hatred and anger it is in your heart .. most people can not do it .. |
| | mm+ft.en | If you say it's hate and angry in your heart, right... when you smoke the trigger, most people can't do it. |
| | mm.en | If it's a hate and anger in your heart, then it's okay... when it's time to smoke the treasure, most people can't do it. |
| WB-2 | wat.ml | വാളിനേക്കാൾ മൂർച്ച് അതു ചലിപ്പിക്കുന്നവൻറെ ചിന്തയിലാണ് ഉണ്ടാകേണ്ടത്. |
| | wat.en | Sometimes a sharp mind is enough. |
| | google.en | It should be in the mind of the one who moves it over the sword and moves it. |
| | mm+ft.en | It's more important than the sword that's in the mind of a moving man. |
| | mm.en | It's supposed to be in the mind of the sword that drives. |
| WB-3 | wat.en | "Most likely they have died of cold and hunger - far away there in the middle of the forest." |
| | wat.ml | "Most likely they have died of cold and hunger - far away there in the middle of the forest." |
| | mm.ml | "അവർക്ക് തണുപ്പും വിശപ്പും മരണം സംഭവിക്കും... ...അവിടെ വനത്തിൻറെ മദ്ധ്യത്തിൽ ദൂരെ." |
| | mm+ft.ml | "അവർക്ക് തണുത്തതും വിശപ്പും മൂലം മരണം സംഭവിക്കാം... ...അവിടെ കാട്ടിൻറെ നടുവിൽ ദൂരെ." |
| WB-4 | wat.en | "Most likely they have died of cold and hunger - far away there in the middle of the forest." |
| | wat.en | He told me you would get better. And you did. |
| | wat.ml | അവൻ പറഞ്ഞു അമ്മയ്ക്ക് വേഗം ഭേദമാകുമെന്ന് അതു മാറുകയും ചെയ്തല്ലോ. |
| | mm.ml | അവൻ എന്നോട് പറഞ്ഞു നിനക്ക് നല്ലത് കിട്ടുമെന്ന്. |
| | mm+ft.ml | അവൻ എന്നോട് പറഞ്ഞു നിനക്ക് സുഖമാവുമെന്ന്. |
| WB-5 | wat.en | Old age should burn and rave at close of day |
| | wat.ta | Old age should burn and rave at close of day |
| | mm.ta | வயதானவர்கள் நாள் முழுவதும் எரிக்க வேண்டும்... |
| | mm+ft.ta | பழைய வயது எரிக்க வேண்டும் மற்றும் நாள் நெருக்கமாக குழாய் வேண்டும் |

**Table 4.6** Failure Cases from WAT test-set

| | |
|---|---|
| source | Or you could leave and return to your families as men instead of murderers. |
| target | അല്ലെങ്കിൽ നിങ്ങൾക്ക് കൊലയാളികൾക്ക് പകരം നിൻറെ കുടുംബത്തിൽ മടങ്ങിപ്പോകാൻ കഴിയും. |
| source | I always wanted to build a talkies in our village in your father's name. |
| target | നിന്റെ അച്ഛന്റെ പേരിൽ നാട്ടിൽ ഒരു ടാക്കീസ് പണിയുക എന്നത് പണ്ട് മുതലേ ഉള്ള ഒരഗ്രഹമാണ്, മോനെ. |
| source | At the funfair, near the ghost train, the marshmallow twister is twisting. |
| target | मेले में, भूतीया train के पास, marshmallow घूमता ज रह है |
| source | See how lovingly he still stares at his master? |
| target | देखो कि वह अपने गुरु पर अब भी कैसे प्यार करता है? |
| source | And that's for Vortex |
| target | এবং সেটা কিনা ভরটেক্স'র হয়ে? পরিষ্কার বুঝা যাচ্ছে তোমার কাছে সবার নাম আছে। |

**Table 4.7** Qualitative sample from WAT test-dataset. Missing tokens are colored in red. Few samples which are treated as gold aren't as good.

*MULTI+ft* . In this section, we present supplementary content assessing *IL-MULTI+ft* , specific to the merits in particular to translation tasks WAT-ILMPC.

Tables 4.5 and 4.6, we contains success and failure cases of our models on WAT-ILMPC test-sets, already discussed before.

### 4.3.2 Mann Ki Baat

In the previous section, we examined the results qualitatively to understand the performance metrics that come along with it. However, due to problems with the test-corpus and the nature of the test-corpus being English-centric, leads to the necessity of other sources for a multiway test-set. In this direction, we present our multilingual translation BLEU scores on *Mann Ki Baat* test-set.

We find the performance of the model is what is expected in agreement with existing NMT literature. Our baseline model gives us good results on the *Mann Ki Baat* test set. However, contrary to the case in WAT-ILMPC, we find that *IL-MULTI* outperforms *IL-MULTI+ft* . This is expected as we fine-tuned to a subtitles corpus, and got better off results. As widely demonstrated in literature, augmentation with backtranslated data gives improvements of few BLEU points to models. Overall *IL-MULTI+bt* performs the best, as visible from the grids given below (Tables 4.9, 4.8, 4.10).[2]

[2]We are following indicnlp's BLEU computation, which should match the WAT website's BLEU computation, but did not. We are in touch with the organizers to make the values consistent.

| srcs | bn | en | hi | ml | ta | te | ur |
|---|---|---|---|---|---|---|---|
| bn | 99.49 | 9.53 | 3.07 | 1.07 | 0.59 | 1.46 | 0.00 |
| en | 4.19 | 99.99 | 10.10 | 1.93 | 1.18 | 1.63 | 0.00 |
| hi | 4.54 | 13.28 | 99.42 | 1.50 | 0.91 | 1.36 | 18.44 |
| ml | 1.33 | 7.51 | 4.79 | 99.42 | 0.88 | 1.89 | 0.00 |
| ta | 0.98 | 4.24 | 3.21 | 0.92 | 99.53 | 0.93 | 0.00 |
| te | 2.06 | 5.79 | 3.15 | 1.41 | 0.47 | 99.10 | 0.00 |
| ur | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100.00 |

**Table 4.8** BLEU scores of *IL-MULTI+ft*

| srcs | bn | en | hi | ml | ta | te | ur |
|---|---|---|---|---|---|---|---|
| bn | 99.85 | **11.56** | 12.42 | 2.01 | 1.20 | 2.30 | 0.00 |
| en | 4.98 | 100.00 | **13.75** | 1.87 | 1.28 | 1.81 | 18.52 |
| hi | 6.93 | 17.87 | 99.49 | 2.77 | 1.71 | 3.09 | 12.55 |
| ml | 3.19 | **8.93** | 8.79 | 99.77 | 0.95 | 2.43 | 0.00 |
| ta | 1.68 | 5.15 | 5.33 | 1.16 | 99.74 | 1.57 | 0.00 |
| te | 2.78 | 7.51 | 7.67 | 2.24 | 1.06 | 99.48 | 0.00 |
| ur | 0.00 | 0.00 | 32.64 | 0.00 | 0.00 | 0.00 | 100.00 |

**Table 4.9** BLEU scores of *IL-MULTI*

| srcs | bn | en | hi | ml | ta | te | ur |
|---|---|---|---|---|---|---|---|
| bn | 99.85 | 11.28 | **12.46** | **2.49** | **1.30** | **2.72** | 0.00 |
| en | 5.18 | 100.00 | 13.66 | **2.78** | **1.46** | **2.81** | 0.00 |
| hi | **7.01** | **17.95** | 99.50 | **3.03** | **2.07** | **3.78** | 0.00 |
| ml | **3.45** | 8.86 | **8.98** | 99.77 | **1.04** | **3.08** | 0.00 |
| ta | **1.78** | **5.28** | **5.46** | **1.22** | 99.85 | **1.71** | 0.00 |
| te | **3.01** | **7.64** | **7.84** | **2.61** | **1.21** | 99.68 | 0.00 |
| ur | 0.00 | 0.00 | 32.64 | 0.00 | 0.00 | 0.00 | 100.00 |

**Table 4.10** BLEU scores of *IL-MULTI+bt*

| | | |
|---|---|---|
| en | Mann Ki Baat, February 2019 My dear countrymen, Namaskar. | |
| en | Mann Ki Baat, February 2019 My dear countrymen, Namaskar. | Mann Ki Baat, February 2019 My dear countrymen, Namaskar. |
| hi | मान की बात, फरवरी 2019 मेरे प्रिय देशवासी, नमस्कार। | मन की बात, फरवरी 2019 मेरे प्यारे देशवासियो, नमस्कार |
| bn | কি বাট, ফেব্রুয়ারি ২০১৯ আমার প্রিয় দেশবাসী, নামসকার. | মন কি বাত, ফেব্রুয়ারি ২০১৯ আমার প্রিয় দেশবাসী, নমস্কার! |
| ta | மன் கி பாட், பிப்ரவரி 2019 என் டியர் நேச நாட்டு நமஸ்கர். | மனதின் குரல், பிப்ரவரி 2019 எனதருமை நாட்டு-மக்களே, வணக்கம்! |
| en | In order to ensure that their fellow countrymen could sleep peacefully, these brave sons toiled relentlessly, day or night. | |
| en | In order to ensure that their fellow countrymen could sleep peacefully, these brave sons toiled relentlessly, day or night. | In order to ensure that their fellow countrymen could sleep peacefully, these brave sons toiled relentlessly, day or night. |
| ml | ദേശവാസികൾക്ക് സമാധാനത്തോടെ ഉറങ്ങാൻ കഴിയുമെന്ന് ഉറപ്പുവരുത്താൻ, ഈ ധീരപുത്രൻമാർക്ക് നിത്യമോ, രാത്രിയോ ഉണർന്നില്ല. | ജനങ്ങൾ സമാധാനത്തോടെ ഉറങ്ങാൻ, നമ്മുടെ ഈ വീരപുത്രന്മാർ രാത്രിയെ പകലാക്കി കാവൽ നിന്നു. |
| hi | ताकि यह सुनिश्चित किया जा सके कि उनके साथी देशवासी शांतिपूर्वक सो सकें, इन बहादुर बेटों ने अथक, दिन या रात बिता दी। | देशवासी चैन की नींद सो सकें, इसलिए, इन हमारे वीर सपूतों ने, रात-दिन एक करके रखा था |
| te | దేశస్థులు ప్రశాంతంగా నిద్రపోవాలని శ్రద్ధగా నిశ్చయించేందుకు ఈ సాహస కుమారులు ప్రశాంతంగా, పగలు లేదా రాత్రికి తోడ్పడ్డారు. | దేశప్రజలు ప్రశాంతంగా నిద్ర పోవడం కోసం ఈ వీరపుత్రులు తమ నిద్రాహారాలు మానుకుని మనల్ని రక్షించారు. |
| ta | நண்பர்கள் அமைதியாக தூங்க முடிந்தது என்பதை உறுதி செய்ய இந்த வீரப் மகன்கள் அமைதியாக, நாள் அல்லது இரவில் துறந்தனர். | நாட்டுமக்கள் நிம்மதியாக உறங்கவேண்டும் என்பதற்காக, நம்முடைய இந்த வீர மைந்தர்கள், இரவு பகல் எனப்பாராமல் தங்களை அர்ப்பணித்திருந்தார்கள். |
| bn | তাদের দেশবাসীরা শান্তিতে ঘুমাতে পারতে নিশ্চিত করার জন্য, এই সাহসী ছেলেরা নিঃসন্দেহে, দিন বা রাতে বিরক্ত হয়ে যায়। | দেশবাসী যাতে নিশ্চিন্ত ঘুমোতে পারেন সেদিকে খেয়াল রেখে এইসব বীর সুসন্তানরা দিন-রাত এক করে সজাগ দৃষ্টি রেখে চলতো। |

**Table 4.11** Qualitative samples from Mann Ki Baat Dataset. Source is one large row. Translations to different languages are shown in two columns - first one is our prediction and second one the respective aligned ground truth.

| | |
|---|---|
| en | Jamuna Tudu, famous nicknamed ’Lady Tarzan’ in Jharkhand, most valiantly took on the Timber Mafia and Naxalites, and not only saved the 50 hectares of forest but also inspired ten thousand women to unite and protect the trees and wildlife. |

| | | |
|---|---|---|
| en | Jamuna Tudu, famous nicknamed ’Lady Tarzan’ in Jharkhand, most valiantly took on the Timber Mafia and Naxalites, and not only saved the 50 hectares of forest but also inspired ten thousand women to unite and protect the trees and wildlife. | Jamuna Tudu, famous nicknamed ’Lady Tarzan’ in Jharkhand, most valiantly took on the Timber Mafia and Naxalites, and not only saved the 50 hectares of forest but also inspired ten thousand women to unite and protect the trees and wildlife. |
| te | జార్ఖండ్‌లో ప్రసిద్ధి చెందిన "లాడీ టార్జాన్" పేరుగల "జమునా టూడూ", అత్యధికంగా టింబర్ మాఫియా మరియు క్షిపణిదారులపై తీసుకున్నారు, కేవలం 50 హెక్టార్ల అడవులను మాత్రమే కాపాడలేదు, కానీ పదివేల మంది స్త్రీలు చెట్లను మరియు వన్యప్రాణిని సంరక్షించడానికి కూడా ప్రేరణ చేశారు. | ార్ఖండ్ లో లేడీ టార్జాన్ పేరుతో ప్రఖ్యాతి చెందిన జమునా టూడూ, టింబర్ మాఫియా తోనూ, నక్సలైట్ల తోనూ పోరాడే సాహసవంతమైన పని చేసారు. ఏభై హెక్టార్ల అటవీ ప్రాంతాన్ని నష్టపోకుండా కాపాడడమే కాకుండా, కలిసికట్టుగా చెట్లు, వన్యజీవాల రక్షణ కోసం పోరాడేలా పదివేల మహిళలకు ప్రేరణను ఇచ్చారు. |
| ta | **ஜார்க்கண்டில் புகழ்பெற்ற "லாடி தர்ஜன்" ("Jamuna Tudu", "Timber Mafia" மற்றும் "Naki") மீது மிகவும் பிரபலமாக எடுத்தார், மேலும் 50 ஹெக்டேர் காடுகளை மட்டுமே காப்பாற்றியதுடன், பத்து ஆயிரம் பெண்களை மரங்களையும், வனவிலங்குகளையும் ஒருமைப்படுத்தவும் தூண்டினார்.** | **நூறு ஆண்டுகள் ஆன போதிலும் கூட உலகம் முழுவதிலும் மக்களுக்கு யோகக்கலையை அவர் பயிற்றுவிக்கிறார், இதுவரை 1500 பேர்களை யோகப் பயிற்றுநர்களாக ஆக்கியிருக்கிறார்.** |
| ml | ഝാർഖണ്ഡിലെ 'ലാഡി ടാർസൻ' എന്ന പ്രശസ്തനായ ജമുന ടുഡു, ഏറ്റവും വിശാലമായി ടിംബർ മാഫിയയും ദേശാടനക്കാരും സഞ്ചരിച്ച് 50 ഹെക്ടർ വനം രക്ഷിച്ചു മാത്രമല്ല, പത്ത് ആയിരം സ്ത്രീകൾ മരങ്ങളും വന്യജീവികളെയും ഒന്നിച്ചു രക്ഷിക്കാനും പ്രേരിപ്പിച്ചു. | ാർഖണ്ഡിൽ ലേഡി ടാർസൻ എന്ന പേരിൽ വിഖ്യാതയായ യമുനാ ടുഡു തടി മാഫിയയോടും നക്സലുകളോടും പോരാടുകയെന്ന സാഹസം പ്രവർത്തിച്ചു. 50 ഹെക്ടർ വനം നശിക്കാതെ കാത്തുവെന്ന മാത്രമല്ല, പതിനായിരം സ്ത്രീകളെ സംഘടിപ്പിച്ച് വൃക്ഷങ്ങളുടെയും വന്യജീവികളുടെയും രക്ഷക്കായി പ്രവർത്തിക്കാൻ അവരെ പ്രേരിപ്പിച്ചു. |
| bn | ঝাড়খন্ডে বিখ্যাত "লাডি টার্জন", সবচেয়ে চম□কারভাবে টিম্বার মাফিয়া ও নক্ষত্রবাহিনী দখল করেন। শুধুমাত্র ৫০ হেক্টর বন রক্ষা করেন, এমনকি দশ হাজার মহিলাকে গাছ ও বন্য প্রাণীকে একত্র করার জন্যও প্রেরণ করেছেন।. | ঝাড়খন্ডের বিখ্যাত □Lady Tarzan□ হিসেবে খ্যাত যমুনা টুডু, গাছপাচারকারী মাফিয়া ও নকশালদের মোকাবিলা করার মত সাহসী কাজ করে তিনি শুধু ৫০ হেক্টর জঙ্গল উজাড় হয়ে যাওয়ার হাত থেকে বাঁচাননি, উপরন্তু দশ হাজার মহিলাকে একজোট করে গাছ ও বন্যপ্রাণীদের সুরক্ষার জন্য পাঠিয়েছেন। |
| hi | झारखंड में मशहूर उपनाम 'लेडी तर्जन' जमुना तुडू ने सबसे ज्यादा चमत्कार से टिंबर माफिया और नक्सलियों को न केवल 50 हेक्टेयर जंगल से बचाया बल्कि दस हजार महिलाओं को पेड़-पौधों और वन्यजीवों की सुरक्षा के लिए प्रेरित किया। | झारखण्ड में □□□ □□□□□ के नाम से विख्यात जमुना टुडू ने टिम्बर माफिया और नक्सलियों से लोहा लेने का साहसिक काम किया उन्होंने न केवल 50 हेक्टेयर जंगल को उजड़ने से बचाया बल्कि दस हज़ार महिलाओं को एकजुट कर पेड़ों और वन्यजीवों की सुरक्षा के लिए प्रेरित किया |

**Table 4.12** Longer sentences from Mann Ki Baat set.

## 4.4 Revisiting translation in narrow domain

Transfer learning has been widely used by the computer vision community in the past. These works typically achieve excellent performance on a specific narrow domain task by fine-tuning a generic model trained on a larger dataset like ImageNet [21]. Similarly in the language space, past works attempt at fine tuning embeddings for sentiment classification to a new language or domain where data is scarce, effectively transferring learning from the general case the embeddings are originally trained on. In this section, we experiment similar fine-tuning methods and show that we can improve our translation system for a particular domain – cricket related content.

Newspaper reports and commentary archives on cricket are crawled and corrected, and a noisy parallel pair is obtained using online services. 279K parallel pairs are thus obtained between Hindi and English. We set aside 100 samples as test-set and collect human provided translations for the same.

| Model | en → hi | hi → en |
|---|---|---|
| cold-start | 30.80 | 33.84 |
| warm-start | 25.57 | 29.27 |
| + fine-tuning | **35.72** | **39.97** |

**Table 4.13** BLEU scores of different takes on building a domain specific translation system, for cricket. Cold-start denotes model trained on cricket data alone. Warm start begin with *IL-MULTI* - trained on the general dataset. This model is further fine-tuned to cricket data.

We use three types of models while testing - (1) *IL-MULTI* , (2) *IL-MULTI* fine-tuned on cricket data and (3) a model with same parameters as *IL-MULTI* trained from scratch. Table 4.13 compares the performance of the three models on the test set. Among the three models, the warm-started model fine-tuned on cricket specific corpus achieves the best performance, outperforming the model trained from scratch on this narrow domain.

Cricket domain include many terms which resolve to imply meanings different from their normal ones, when restricted to the domain. For instance a “boundary” or “four” would mean the event of “four runs” being scored in a match. Terms like “sixers”, “square drive”, “form” etc pose different meanings compared to the meanings associated with the same when placed in a general context. In Table **??** are some examples corrected by the fine-tuning method, which would otherwise be mistranslations due to the lack of contextual knowledge.

## 4.5 Conclusions

In this chapter, we have demonstrated improvements in NMT performance in directions involving 7 Indian Languages and English. We take the first step towards a multilingual test-set across several Indian

languages, enabling study and comparisons of multilingual models which operate in these languages. In addition we revisit the narrow domain work to demonstrate improvements through domain adaptation by fine-tuning a general-purpose model.

Our investigations with NMT involving more languages pointed to the need of better-datasets. Enhancing the available data across language pairs is further supported by the increase in digital content online, with more of the countries population gaining access to the online content. In the next chapter, we discuss such an effort involving the Press Information Bureau, which we identified as a source which continuously improves with more news articles added. In addition, we update and refine the Mann Ki Baat test-set to make it better for future translation benchmarks in the language-directions.

# Chapter *5*

# Applications

In this chapter, we describe methods using the Multilingual NMT systems developed in the past chapters to mitigate the data scarcity in Indian languages. We identify multilingual models and parallel pairs aligned across multiple languages as a promising frontier.

## 5.1 Enhancing translation resources for Indian Languages

### 5.1.1 Motivation

Modern day neural network based approaches for Machine Translation (MT) are data hungry and sentence-level aligned parallel pairs are the currency. Neural MT (NMT) is currently the de-facto approach for for training translation systems and shares the same traits [43].

Koehn [41] uses European parliament proceedings available on the web to create an evolving parallel corpora. The resource has been a major driving factor in attempts to build MT systems. We provide a similar effort through this work in providing a multilingual parallel corpora for Indian languages.

Many languages spoken in the Indian subcontinent are categorized as low-resource [77] considering the amount of parallel corpora available for training with deep neural network based models. There have been a few efforts towards addressing this lacunae. Among this, the IIT-Bombay Hindi English Parallel Corpus (IITB-hi-en) [51] is the largest English-Hindi corpus available for training. Other significant efforts in constructing parallel corpus for Indian languages include Indian Language Corpora Initiative (ILCI) [35] across 10 Indian languages, restricted to health and tourism domains and Indic Multilingual Parallel Corpus (WAT-ILMPC) [62], with possible shared content among multiple languages collected through automated efforts are limited and noisy [74]. WAT-ILMPC is comprised of user contributed translations of subtitles, which are inadequate and often code-mixed. A specific effort is that of the Oriya-English Corpus (OdEnCorp) that provides a compilation in Oriya-English [71]. Lack of resources reflects in the availability of automated translation systems of satisfactory quality in these languages. While some systems are reported [99, 74] to perform well for English-Hindi, there are not many that generalize well for other languages. Based on these resources, in order to enable further

progress in Indian languages, additional multi-lingual general purpose corpora that could be used for training over multiple Indian languages can be considered as a required resource. Further, in order to gauge the progress made in solving this task, it is also necessary to consider an independent general purpose test corpus that is not used for training. In this work, we address both these challenges.

Low resource languages have drawn an increased interest towards self-supervised extraction of sentence-aligned bitext [83, 84, 5] to augment training data for MT. Attempts at using pre-training with monolingual corpora and transferring the learning to machine translation have found moderate success [31]. Another class of augmentation approaches [24, 86] turns to using noisy synthetic corpus obtained through back-translation to improve translation results in the low resource setting. However, the more relaxed the formulations are, larger are the computational requirements for some improvement. Moreover, these formulations can be viewed as expectation maximization (EM) algorithms [16, 30] and often benefit from a good initialization - which in the case of MT is a collection of aligned sentences in two languages - a parallel corpus.

Another benefit of obtaining a parallel corpus is that it enables other resources such as cross-lingual word embeddings [81] and sentence embeddings [15] to also be developed. These resources would also enable downstream tasks such as paraphrase generation and question answering in multiple Indian languages that are currently less considered due to lack of sufficient machine translation or word embedding capabilities.

All the above points to a strong need of usable parallel corpora for many languages to catch up. Recent literature suggests multilingual corpora with shared content among languages can provide significant boost to learning [2, 65, 36], enabling a single model to do zero shot translation and acting as regularization. Other neural approaches, which has further relaxed formulations include Schwenk et al. [84]. With many vernacular web sources publishing bilingual content and more speakers gaining access to the internet [74] - the resources available are growing. Combined with recent developments in natural language understanding with advent of deep neural networks - it is possible today to construct corpora with lesser human efforts.

This work relies on two sources – the Press Information Bureau (PIB) and *Mann Ki Baat*, the Indian Prime Minister's speeches to compile sentence aligned parallel corpora with shared content across 10 languages. Our contributions are as follows:

- We revisit MT based alignment methods, to obtain the aligned parallel corpus with minimum human supervision.
- We release a corpus of size 407K compiled across 10 language pairs from PIB, intended as training data for multilingual models.
- We release an independent test corpus compiled from *Mann Ki Baat*. This is a useful resource to validate machine translation accuracy and generalization performance. The test-sets are between 9 languages with 2-3K example sentences on an average.

This chapter is structured as follows: §5.1.2 briefs the method involving NMT based alignment used to obtain sentence pairs. In §5.1.3, we summarize and describe the multilingual parallel corpora, the quality is validated by the evaluations elaborated in §5.1.4 We demonstrate state of the art performance in many tasks using the datasets in §5.1.4

### 5.1.2 Aligned Parallel Corpora from Web

We take advantage of the structured information - links, dates etc. in the above sources of the same while crawling and pre-processing. Once we have alignments between documents, we use several strategies to obtain sentence level alignments across languages. On obtaining sentence level alignments, we use heuristics – length ratios, language identification through writing script etc, based filtering to remove noisy pairs.

**Crawling and Preprocessing** We crawl PIB and *Mann Ki Baat* websites for the articles. We extract only the text from html. Following the text extraction, we index and store the articles on the basis of a unique identification number that is assigned to each article. We additionally store metadata attributes of posted date and language the article was written in. This dataset of documents is subsequently processed to obtain delimited sentences in each document.

For sentence segmentation, we use a rule-based tool created using commonly used sentence delimiters across Indian languages. For Urdu, we use UrduHack[1]. These sentences extracted are used as units for translation for both document and sentence alignments. The sentences are tokenized for the tasks discussed ahead by using SentencePiece [47] models trained by restricting each language to a vocabulary size of 4000 units. The result of this processing leads to a set of documents with delimited sentences in 11 languages for PIB and 10 languages for *Mann Ki Baat* websites. These are then aligned using a document alignment procedure.

**Document Alignments** For obtaining document level alignments in *Mann Ki Baat*, posted date of the article suffices as a retrieval criteria. The articles are posted in periodic manner with the multilingual content being uploaded with negligible time delay. Unlike *Mann Ki Baat*, PIB articles are posted with different timestamps making it harder to retrieve multilingual content.

Given an MT system, we can translate all non-English articles to English. This makes it possible to check similarity between two documents in English - one the original and the other a translation. We use cosine similarity on *term frequency-inverse term frequency* (TF-IDF), a measure commonly used to rank candidates in retrieval literature to obtain candidate articles within the neighbourhood of 2 days of the posted date of source and choose the nearest neighbour to align documents. Using these procedures we obtain a document level alignment between articles in the various languages. From these set of aligned documents, we process and obtain sentence level alignments.

[1] https://github.com/urduhack

**Sentence Alignments** We use BLEUAlign approach [88] where we have a reasonable working translation model between the language pairs. Sennrich and Volk [88] proposed MT based sentence alignment algorithm denoted hereafter by BLEUAlign which uses translation of either source or target text. It uses BLEU score [70] as a similarity metric to obtain sentence level alignments. They demonstrate better alignments than conventional length based alignment methods [25]. The MT systems used in BLEUAlign are discussed next.

| | | en | hi | ta | ur | ml | bn | te | or | gu | pa | mr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PIB | Articles | 28K | 14.5K | 7.5K | 13.5K | 4K | 4.5K | 1K | 6K | 2.2K | 4.7K | 6.4K |
| (train) | Sentences | 1.3M | 440K | 227K | 360K | 97K | 112K | 22K | 89K | 60K | 88K | 165K |
| | Aligned-en | - | 260K | 96K | 122K | 31K | 35K | 10K | 43K | 46K | 61K | 123K |
| | Filtered | - | 156K | 61K | 45.3K | 17K | 21.6K | 6K | 9.1K | 25.5K | 26.3K | 40K |
| | Vocabulary | 159K | 112K | 119K | 67.6K | 67K | 44.2K | 27K | 35K | 51.7K | 53.7K | 73K |
| Mann | Articles | 58 | 57 | 47 | 8 | 45 | 48 | 48 | 39 | 45 | † | 47 |
| Ki Baat | Sentences | 12.2K | 12.6K | 11.6K | 1.7K | 12.3K | 14.7K | 15.2K | 9K | 11.2K | - | 11.9K |
| (test) | Aligned-en | - | 5.4K | 6.0K | 1K | 5.2K | 5.8K | 5.5K | 0.8K | 6.8K | - | 6.2K |
| | Filtered | - | 5.3K | 5.7K | 1K | 5K | 5.6K | 5.2K | 0.8K | 6.6K | - | 5.9K |
| | Vocabulary | 21K | 11K | 23K | 4K | 22K | 17.5K | 21K | 4K | 19K | - | 20K |
| | OOV rate | - | 13.3% | 26.4% | 14.1% | 36.9% | 39.7% | 74.1% | 28.3% | 30.4% | - | 16.8% |

**Table 5.1** Detailed statistics of PIB and Mann Ki Baat. Number of sentences are reported after segmentation of articles. Aligned-en indicates number of sentences aligned to English across each language. Filtered indicates size after filtering aligned sentences through our pipeline. We also report vocabulary compiled from filtered sentences across languages. Details on nature of PIB and Mann Ki Baat corpus in sections 5.1.3.1 and 5.1.3.2 respectively. †Punjabi(pa) is unavailable in Mann Ki Baat.

**NMT for Alignment** We use a single NMT model which is an implementation of the Transformer architecture [98] for a sequence-to-sequence learning task. This architecture consists of an encoder and a decoder. An encoder consumes source sequence represented by one-hot vectors corresponding to the tokens. The decoder uses the encoded representations of the source sequence, attends to it and autoregressively predicts the next token left to right. This constitutes the Transformer architecture. Parameters are shared among all languages for both the encoder and decoder [36]. The embeddings are also shared given the vocabulary at input and output are same through the multilingual formulation and for the benefits reported in Press and Wolf [79]. SentencePiece models help making training multiway models across 11 languages feasible without compromising coverage of all sentences in the language.

| Source | #pairs | type |
|---|---|---|
| IITB-en-hi | 1.5M | en-hi |
| UFAL EnTam | 170K | en-ta |
| Backtranslated-Hindi | 2.5M | en-hi |
| WAT-ILMPC | 800K | xx-en |
| ILCI | 500K | xx-yy |
| OdiEnCorp | 27K | en-or |
| Telugu Bible | 30K | en-te |
| Backtranslated-Telugu | 500K | en-te |

**Table 5.2** Training dataset used for multilingual model. xx-yy indicates parallel sentences aligned across multiple languages.

The dataset used to train our multilingual model is compiled from multiple sources: IIT Bombay English-Hindi corpus [51], UFAL EnTam v2.0 [80], OdiEnCorp [71], ILCI [35] and WAT-ILMPC [62] multilingual corpora. Additionally we compile English-Telugu parallel text from Bible corpus [14]. We also augment training set with synthetic data obtained from Backtranslation [86] for Hindi and Telugu.

**Filtering aligned sentences for noise** Based on the two alignment techniques, we obtain a set of aligned candidate sentence pairs. Among the extracted sentence pairs through the methods in the section above, we observe noisy sentences. These are in the form of URLs, numbers corresponding to dates, segments left untranslated in the news etc. We do not filter these early on as these can help in aligning sentences. However, to ensure that these do not affect training the machine translation model, we filter the data using `langid`[2] to remove foreign language tokens. We also filter based on the ratio of source-to-target length to filter imbalanced sentences amongst the pairs. BLEUAlign also has a filtering effect on sentence pairs as it can merge and discard sentences based on the BLEU score between target sentence and translation of source.

**Compiling Multilingual Content** Finally, to obtain aligned corpora across languages, we group sentence pairs extracted from many languages keyed by a single language. In this work, we use English as the common language. The choice is helped by the fact that many parallel corpora available for Indian languages have English as one of the two languages. Due to this all languages translate to English fluently. Through this procedure we obtain a multi-lingual parallel corpus.

[2] https://github.com/saffsd/langid.py

### 5.1.3 Dataset Description and Statistics

In this section we describe the datasets produced with this work. Two datasets are compiled - the PIB corpora for training and *Mann Ki Baat* corpora for a standard test-set. Both corpora have shared sentences across multiple languages. We elaborate on the nature of the compiled corpora and the quality.

#### 5.1.3.1 Press Information Bureau Corpus

Press Information Bureau (PIB) is an Indian government agency responsible for communications to media. These communications are released in the form of articles on a website[3] across 12 Indian Languages: Hindi (hi), Telugu (te), Tamil (ta), Malayalam (ml), Gujarati (gu), Kannada (kn), Urdu (ur), Bengali (bn), Oriya (or), Marathi (mr), Punjabi (pa), Assamese (as) and English (en). Out of these 12 Indian languages parallel corpora of variable sizes are available for 10 of them, excluding Assamese and Kannada. Articles are manually translated by PIB officials and can be deemed expert translations, which makes this a source for authentic parallel data. Articles are posted for about two years now since 2017. With progress of time more multilingual content will be added by the organization, which warrants improvements in size of corpus across all languages. These translations are adequate and fluent by nature, provided good alignment accuracy is obtainable. We observe that an article on an average provides 10 sentences.

**Corpus Statistics** We propose PIB corpora as a multilingual training set. The corpus compiled out of PIB contains 407K sentences aligned with shared content across multiple languages. More on the corpus statistics are described in Table 5.1. The articles in the corpus are aligned with English as a pivot language. Hindi and Urdu articles are being published since the start. Articles in the other languages are being brought up to speed recently leading to an imbalance between number of articles across languages (14.5K hi to 1K te). We also observe disproportion between number of articles and en-aligned sentences as not all articles posted have correspondence for English article in PIB. This leads to lesser number of aligned documents for extracting sentence pairs. The number of sentence pairs further decreases as BLEUAlign chooses to merge and discard source sentences based on BLEU score between translated source and target pairs. One such example is demonstrated in Table 5.3. Presence of large number of English tokens, numbers and duplicates among the aligned pairs results further reduction in size after filtering across languages. We observe variations in proportion of filtered sentences among languages. The variance in filtered sizes of the corpora is caused by erroneous matches by BLEUAlign which are pruned out by sentence-ratio based thresholding and example of which is present in Table 5.4. A lot many English words are also interspersed in Indian language articles.

**Multilingual shared content** Due to the presence of shared multilingual content among the languages in PIB we were able to extract aligned sentence pairs across multiple languages (xx-yy). The resulting

[3] https://pib.gov.in

corpora is described in Table 5.6 in format of a multilingual grid where we report sizes of sentence pairs aligned across languages. As PIB and *Mann Ki Baat* share all of the languages except one, we merge the obtained multilingual grid with upper triangle color coded Blue representing PIB corpora. en-hi have the highest number of parallel sentences constituting 38% of the corpora followed by en-ta and en-ur at 14% and 11% respectively. Because of these high proportions of en-aligned data we were able to obtain 22.3K hi-mr (related language pair), 26K hi-ta (distant language pair) sentence pairs by pivoting through english (refer Table 5.6). Such alignments among diverse set of languages can significantly boost the performance of multilingual NMT systems.

#### 5.1.3.2 Mann Ki Baat

*Mann Ki Baat*[4] is a transcribed source for Indian Prime Minister's speeches across 12 Indian languages: Hindi, Telugu, Tamil, Malayalam, Gujarati, Kannada, Urdu, Bengali, Oriya, Marathi, Assamese, Manipuri (mp) and English. As the volume of this transcriptions is much less compared to PIB we compile a multilingual test. As these articles are relatively sparse[5], posted date serves as primary criterion for retrieving the multilingual content. These translations can be considered as expert translations and are fluent by nature as the speeches are addressed to people across India speaking many languages. These transcribed speeches usually tend to be longer, unlike PIB articles. *Mann Ki Baat* is also a promising source for multilingual corpora as more content is being added over time.

**Corpus Statistics** We release *Mann Ki Baat* as a multilingual test set upon which our models are evaluated. We extract sentences aligned with English as the pivot language compiling a bi-directional (xx-en) test set, described in Table 5.1. Volume of Urdu articles posted in *Mann Ki Baat* is relatively less resulting in fewer aligned sentences. Noise in the form of untranslated tokens and repetitions is also present in sentence aligned pairs of *Mann Ki Baat* although not as prevalent as in PIB resulting in removal of only a fraction of sentences after filtering. As *Mann Ki Baat* is released as a test set we compute OOV (Out-of-Vocabulary) rate for word types between PIB and *Mann Ki Baat* corpora (refer Table 5.1). We observe higher OOV rate among te (74.1%), bn (39.7%) as the size of English aligned corpus is less for these languages compared to others in PIB.

**Multilingual shared content** *Mann Ki Baat*, similar to PIB has shared multilingual content across languages. We extract sentence pairs (xx-yy). The resulting corpora is described in Table 5.6 in the format of a multilingual grid. Bottom triangle of the grid color coded in Red represents *Mann Ki Baat*. Presence of less number of en-aligned pairs among Urdu (ur) and Oriya (or) consequently led to less number of pairs across the grid for these two languages. For other languages we obtain test sets with sentences greater than 5K leading to significant number of aligned pairs (3K on an average) across all the

[4] https://www.narendramodi.in/mann-ki-baat/

[5] Prime Minister's speeches happen in a periodic manner, usually with frequency of 1 per month.

languages in the grid. Test corpora of this size aligned across multiple language is previously unavailable, and can benefit standardization of multilingual models built for Indian languages.

#### 5.1.3.3 Characterizations and Comparisons

Multilingual corpora available currently for Indian languages are ILCI [35] and WAT-ILMPC [62]. In this section we compare PIB and *Mann Ki Baat* with these multilingual corpus.

ILCI is a multilingual corpora with 50K sentence aligned pairs across 11 Indian languages and English and has marked a huge contribution in developing Translation systems for Indian languages since its release in 2010. However, ILCI is restricted to only health and tourism domain. In our past experiments, we observe that training on ILCI alone does not generalize well to real-world queries. ILCI dataset is only available for Indian nationals[6]. WAT-ILMPC is a multilingual corpora of subtitles collected from OPUS[7]. This corpus contains lot of short and untranslated segments. As the PIB corpus is translated by experts, we observe a rich vocabulary of words. In Table 5.5, we summarize the characteristics of these alongside PIB for comparison.

Test sets of multilingual nature available for Indian languages currently are aligned only to English [62]. We observe the similar problem of short and untranslated tokens in test set which implies that good performance on this test set does not generalize good performance overall. *Mann Ki Baat* contains sentence alignments across 10 language pairs and English resulting in better understanding of translation phenomena across diverse set of Indian languages, some of them which are related (hi-mr, hi-gu) and some distant (hi-ta). *Mann Ki Baat* are also expert translation, containing a rich vocabulary. This test can be viewed as general domain as these are compiled from transcribed speeches addressing national crowd.

### 5.1.4 Discussions

To demonstrate efficacy of the parallel corpus, we propose two methods of evaluation. First, we report the alignment quality across a small sample reported by humans. Second, we take two cases to check if the parallel corpora created are useful in Machine Translation - we train models one on the PIB corpus alone and also use the same to augment training with existing data and demonstrate improvements in translation quality in the augmented case over the other.

#### 5.1.4.1 Alignment Quality

We randomly sample 100 sentence pairs across all the languages of PIB and *Mann Ki Baat* after aligning to English. We manually evaluate the quality of alignments for these sentence pairs and report the number of sentences that have been verified as correct alignments in Table 5.7. In PIB, we observe

[6] http://www.statmt.org/wmt19/translation-task.html - Gujarati

[7] http://opus.nlpl.eu/

| Index | Corpus | Source | Target |
|---|---|---|---|
| 1 | PIB | इनमें से 3,18,931.22 करोड़ रुपये रक्षा बजट (रक्षा पेंशन के अतिरिक्त) के रूप में हैं। | Out of this Rs.3,18,931.22 crore has been earmarked for Defence (excluding Defence Pension). |
| 2 | PIB | इस अवसर पर विदेश मंत्रालय, सरकारी विभागों के वरिष्ठ अधिकारी, 38 पासपोर्ट अधिकारी और सेवा प्रदाता टीसीएस के अधिकारी उपस्थित थे। सर्वश्रेष्ठ कार्य करने वाले पासपोर्ट अधिकारियों और पुलिस अधिकारियों को बेहतर पासपोर्ट और सत्यापन के लिए उत्कृष्टता पुरस्कार दिए गए। | Senior officials of the ministry of External Affairs and other government departments beside 38 Passport officers and officials of service provider TCS attended the function where awards of excellence were presented to the best performing Passport Officers and Police Officers for best passport and verification services. |
| 3 | Mann Ki Baat | वर्षों पहले डॉ बाबा साहब आम्बेडकर ने भारत के औद्योगिकीकरण की बात कही थी। | Years ago, Dr. Baba Saheb Ambedkar spoke of Indias industrialization. |
| 4 | Mann Ki Baat | मैंने किसी को खादीधारी बनने के लिये नहीं कहा था । लेकिन मुझे खादी भण्डार वालों से जानकारी मिली कि एक सप्ताह में करीब करीब सवा सौ परसेन्ट हंड्रेड एंड ट्वेंटी फाइव परसेन्ट बिक्री में वृद्धि हो गयी । | I had not asked anyone to be Khadidhari, But the feedback I got from Khadi stores was that in a weeks time the sales had jumped up by 125. |

**Table 5.3** Success cases of Alignment for sentence pairs from PIB and Mann Ki Baat. 1 and 3 are shorter sentences but relatively hard to align to English due to presence of sentence delimiters on the English side. 2 and 4 have longer Hindi sentence containing two segments delimited by *Purnavirama* (Hindi end of sentence marker) symbol. We observe that *BLEUAlign* merges these segments to obtain a higher BLEU score, aligning to a longer Hindi (1&3) and English (2&4) sentence which conveys same meaning as the source.

| Index | Corpus | Source | Target |
|---|---|---|---|
| 1 | PIB | वो पीढ़ी दर पीढ़ी तैयार होते जाएंगे उसी क्षमता के साथ, सामर्थ्य के साथ तैयार होंगे। | They will be prepared about every generation. |
| 2 | PIB | रूपया 17 अप्रैल, 2017 को 64.41 रुपये प्रति अमेरिकी डॉलर के स्तर पर बंद हुआ। | 2017. Rupee closed at Rs. 64. |
| 3 | Mann Ki Baat | यहा के प्राथमिक विद्यालय के शिक्षक पी.के. मुरलीधरन और छोटी सी चाय की दुकान चलाने वाले पी.वी. चिन्नाथम्पी, इन दोनों ने, इस ला-इब्रेरी के लिए अथक परिश्रम किया है | V. Chinnathampi who runs a small tea shop, have between them worked tirelessly for this library. |

**Table 5.4** Failure cases of Alignment for sentence pairs from PIB and Mann Ki Baat. 1 represents a hard case to align, where the content is written elaborately in source and poorly in target language. 2 contains English sentence with punctuation, misaligned due to segmentation. 3 represents a case of partial alignment due to segmentation of the English sentence.

| Source | #sents | Vocab | #langs | Languages (xx) |
|---|---|---|---|---|
| ILCI | 550K | 640K | 11 | bn en hi ml ta te mr gu ur ka pa |
| WAT-ILMPC | 800K | 780K | 8 | bn en hi ml si ta te ur |
| PIB | 407K | 780K | 11 | bn en hi ml ta te mr gu ur pa or |
| *Mann Ki Baat* | 41K | 154K | 10 | bn en hi ml ta te mr gu ur or |

**Table 5.5** Multilingual datasets available for Indian languages.

| | en | hi | ta | te | ml | ur | bn | gu | mr | or | pa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| en | | 156344 | 60836 | 6035 | 17187 | 45355 | 21615 | 25598 | 40200 | 9101 | 26338 |
| hi | 5272 | | 25848 | 1845 | 7865 | 9635 | 10476 | 13243 | 22378 | 2307 | 7751 |
| ta | 5744 | 2761 | | 2275 | 5327 | 3416 | 6704 | 8575 | 12180 | 974 | 4778 |
| te | 5177 | 2289 | 3100 | | 1196 | 564 | 917 | 1909 | 829 | 183 | 1086 |
| ml | 5017 | 2305 | 3124 | 2898 | | 1220 | 3243 | 4309 | 3628 | 376 | 1921 |
| ur | 1019 | 742 | 637 | 599 | 624 | | 1631 | 2100 | 1500 | 741 | 5214 |
| bn | 5634 | 2706 | 3460 | 2939 | 2938 | 559 | | 4718 | 4618 | 490 | 1906 |
| gu | 6615 | 3213 | 3998 | 3528 | 3469 | 749 | 3810 | | 5580 | 620 | 3702 |
| mr | 5867 | 2491 | 3175 | 2839 | 2803 | 490 | 3054 | 3658 | | 867 | 3080 |
| or | 768 | 389 | 470 | 440 | 427 | 98 | 447 | 541 | 432 | | 889 |
| pa | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

**Table 5.6** Multilingual shared content across language pairs for PIB (Blue) and Mann Ki Baat (Red). Rows and columns indicate language pairs. Upper triangle (Blue) indicates PIB across Languages. Lower triangle (Red) indicates Mann Ki Baat. The intensity of the cell color is proportional to size of the sentences aligned pairs. † Mann Ki Baat does not contain Punjabi.

high percentages for Urdu and Telugu indicating a good quality of alignment. *Mann Ki Baat* exhibits higher alignments accuracies due to a very strict one-to-one mapping between sentences in the speeches. On the other hand, PIB articles have relaxed writing, often merging or splitting sentences written in one language going to to the other.

| Language Pairs | PIB | Mann Ki Baat |
|---|---|---|
| en-hi | 94% | 99% |
| en-ta | 94% | 98% |
| en-te | 97% | 100% |
| en-ml | 93% | 100% |
| en-ur | 96% | 100% |
| en-bn | 87% | 99% |
| en-mr | 87% | 99% |
| en-or | 2% | 95% |
| en-gu | 91% | 100% |
| en-pa | 90% | N/A |

**Table 5.7** Assessment of sentence alignment quality.

In Oriya, our translation models do not perform as well, evident from NMT performance reported in Table 5.8 which consequently leads to poor retrieval based alignments. We could not observe many articles we could align with pure date based matches either. Mismatch in document alignments further cause noisy sentence alignments for the case of PIB. With the progress of time more data will be made available on the PIB website leading to better mapping of Oriya articles and subsequently better alignments. This will improve the performance of NMT models. In current state, we provide the corpus except for Oriya where the alignment accuracy is observed to be reasonably good. However, for *Mann Ki Baat*, we observe that despite the same model, sentences are aligned well. The articles matching correctly and the existence of a strict one to one mapping in the sentences significantly helps in this case.

Qualitative examples in English-Hindi of success and failure of alignment pipeline are presented in Tables 5.3 and 5.4. We are able to align and extract even significantly long and complex sentences from PIB and Mann Ki Baat. Among the failure cases in Table 5.4, where a fragment matches to a larger sentence due to and alignment error, the possiblity to bring down errors by heuristics like length ratio can be observed.

#### 5.1.4.2 Translation Quality

In order to evaluate translation quality, we train a standard NMT system [74] across all language pairs in a multilingual NMT. These are trained for various cases such as benchmarking training with ILCI or

| | ILCI | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | en | hi | ta | te | ml | ur | bn | gu | mr | pa |
| en | 0.0 | 30.0 | 5.17 | 9.18 | 5.59 | 23.5 | 14.3 | 21.8 | 13.6 | 22.7 |
| hi | 29.0 | 0.0 | 8.12 | 16.1 | 8.78 | 58.6 | 21.3 | 48.3 | 26.7 | 54.0 |
| ta | 13.6 | 19.2 | 0.0 | 7.43 | 5.36 | 16.5 | 9.72 | 15.0 | 9.23 | 16.3 |
| te | 19.8 | 28.7 | 5.62 | 0.0 | 6.61 | 24.0 | 13.6 | 22.8 | 13.8 | 22.0 |
| ml | 17.0 | 23.3 | 5.78 | 10.0 | 0.0 | 19.4 | 13.0 | 18.9 | 11.3 | 18.9 |
| ur | 26.0 | 58.9 | 6.65 | 13.6 | 8.04 | 0.0 | 17.3 | 39.0 | 20.2 | 41.5 |
| bn | 20.8 | 31.3 | 5.83 | 11.5 | 7.21 | 24.5 | 0.0 | 24.2 | 14.7 | 22.8 |
| gu | 27.8 | 57.3 | 7.35 | 15.1 | 8.67 | 45.4 | 20.3 | 0.0 | 23.8 | 40.9 |
| mr | 24.0 | 43.8 | 6.7 | 13.2 | 8.09 | 33.9 | 17.8 | 33.0 | 0.0 | 32.1 |
| pa | 29.2 | 70.8 | 7.51 | 15.0 | 9.13 | 55.3 | 20.0 | 45.4 | 24.5 | 0.0 |

**Table 5.8** BLEU scores of model trained on unaugmented data on multilingual test sets - test-split sampled from ILCI from the unaugmented model. Rows correspond to source languages and columns target languages.

with PIB. We use Bilingual Evaluation Under Study (BLEU) [70] a modified precision based score used in sequence-to-sequence tasks to compare the models trained by different variations of data. We report the BLEU scores obtained using Indic NLP Library[8]. We first consider the multilingual model across languages that we use. We next consider the utility of the proposed PIB corpora. In all our evaluations we also consider evaluating the model with the proposed new *Mann Ki Baat* test set.

**Multilingual scores** Multilingual models enable zero-shot translation [36]. They can translate among unseen pairs of languages during inference which are unseen at training time by implicitly pivoting through shared content. This enables reporting BLEU scores across languages in multilingual grid format in Table 5.8, given the presence of a test set - *Mann Ki Baat* with samples in all directions. Similarly, we report the scores on ILCI with a random subset held out from the training set. Training of all the languages is done on the specified corpora and test is done on the ILCI and *Mann Ki Baat* test set.

#### 5.1.4.3 PIB for NMT

In order to check the usefulness of PIB corpora, we compute the BLEU scores on data trained with all the data we have available and the same data augmented with PIB corpora our NMT model. We

[8] https://anoopkunchukuttan.github.io/indic_nlp_library/

| | Mann Ki Baat | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | en | hi | ta | te | ml | ur | bn | gu | mr | or |
| en | 0.0 | 13.8 | 2.44 | 2.94 | 2.51 | 8.08 | 4.63 | 7.23 | 4.8 | 0.76 |
| hi | 17.8 | 0.0 | 2.81 | 4.82 | 4.02 | 41.22 | 7.62 | 24.3 | 10.4 | 0.7 |
| ta | 9.7 | 11.1 | 0.0 | 3.41 | 2.95 | 10.14 | 3.23 | 5.61 | 3.68 | 0.49 |
| te | 7.19 | 11.4 | 2.18 | 0.0 | 3.22 | 10.65 | 3.58 | 6.13 | 4.14 | 0.3 |
| ml | 11.2 | 14.5 | 2.43 | 4.25 | 0.0 | 12.31 | 4.24 | 7.36 | 4.72 | 0.31 |
| ur | 15.1 | 39.6 | 1.95 | 3.89 | 3.27 | 0.0 | 6.06 | 18.4 | 7.7 | 0.41 |
| bn | 10.8 | 16.2 | 1.82 | 3.51 | 2.76 | 14.1 | 0.0 | 8.92 | 5.13 | 0.41 |
| gu | 14.6 | 37.3 | 2.26 | 4.55 | 3.61 | 29.52 | 7.02 | 0.0 | 9.73 | 0.5 |
| mr | 11.5 | 19.5 | 2.11 | 3.59 | 3.33 | 17.94 | 5.08 | 12.7 | 0.0 | 0.53 |
| or | 3.12 | 8.41 | 1.38 | 1.46 | 1.15 | 8.89 | 3.19 | 4.89 | 3.11 | 0.0 |

**Table 5.9** BLEU scores of model trained on unaugmented data on multilingual test sets - test-split sampled from Mann-Ki Baat obtained from the unaugmented model. Rows correspond to source languages and columns target languages.

further consider the cross-domain generalization ability of models by training individual multilingual models and assess it on existing WAT-ILMPC, ILCI and the proposed new *Mann Ki Baat* corpora. Due to unavailability of separate evaluation set in ILCI we randomly sample a test set containing 500 sentences.

As can be observed from Table 5.8, training on unaugmented corpora alone results in a reasonable performance on ILCI test but a deficient performance on the new *Mann Ki Baat* test set. This indicates that the *Mann Ki Baat* test set is more challenging as compared to the ILCI test set. Further, the performance of this model on these test are not as good as they are being evaluated in a cross-domain setting.

When we consider the performance of using PIB corpora as an additional corpora used along with unaugmented model, the results are significantly improved. The results for the same are provided in Table 5.11. As can be observed, the results in this case show significant improvement in BLEU scores in most cases. The only language in which we see a small decrease in this setting is in Oriya for *Mann Ki Baat* corpora. This we believe is due to the relatively few sentences obtained for Oriya and the fact that the baseline NMT system for Oriya language is noisy.

Table 5.8 additionally provides insights into the effectiveness of multilingual model trained on comparable languages. Translation across strongly related languages are higher, as the mapping is easier to learn. An example we observe is Hindi, Gujarati and Marathi that belong to Indo-Aryan family[9]. Pairs

[9] https://en.wikipedia.org/wiki/Indo-Aryan_languages

| | ILCI | WAT-ILMPC | Mann Ki Baat |
|---|---|---|---|
| ILCI | 21.7 | 6.15 | 6.52 |
| WAT-ILMPC | 5.98 | 25.4 | 5.1 |
| All data | 30.0 | 20.6 | 13.8 |
| All data + PIB | 31.8 | 21.2 | 15.9 |

**Table 5.10** Cross domain inference. Rows indicate the corpus on which model is trained and columns the test-corpus. BLEU scores are reported for en-hi. All data is listed in Table 5.2

among these languages have high BLEU scores. We also observe high BLEU scores among Hindi, Urdu as it is generally regarded that Urdu and Hindi are quite similar phonetically only differing in writing script[10]. This shared linguistic aspects are very helpful as improving translation numbers in one language in the family helps improving across the family due to multilingual models.

We finally consider a general cross-domain generalization performance of the training across the various corpora. The results for the same are provided in Table 5.10. Among these, the ILCI corpora that has a balanced distribution of sentences across languages performs better. However, the case when we add the proposed new PIB corpora results in a increase in performance across different test sets – ILCI, WAT-ILMPC and Mann Ki Baat. A signifant +2 BLEU point increase can be observed in *Mann Ki Baat*, without compromising on other test-sets.

### 5.1.5 Conclusion and Future Work

In this chapter we have presented parallel corpora for 10 Indian languages and English and demonstrated its effectiveness in multilingual machine translation. Our systems are set to collect more sentence pairs as the more articles are published, similar to Europarl [41]. The nature of the corpora opens avenues to explore on how to efficiently use weakly aligned formulations to improve the corpora iteratively.

Many languages that are related have better translation performance between them compared to English, examples being Hindi-Punjabi, Hindi-Marathi, Hindi-Urdu. MT based alignments between these can be used to create larger mappings amongst the sentences in both. Iterative application of such methods can lead to a very refined corpus, through minimum human annotation.

Finally, we also provide a general test corpora that can be used to independently evaluate MT performance for Indian Languages in a generic setting.

[10] https://www.britannica.com/topic/Urdu-language

Mann Ki Baat

| | en | hi | ta | te | ml | ur | bn | gu | mr | or |
|---|---|---|---|---|---|---|---|---|---|---|
| en | 0.00 | 2.10 | 1.60 | 0.85 | 2.10 | 9.42 | 2.76 | 3.77 | 3.20 | -0.53 |
| hi | 2.40 | 0.00 | 2.74 | 0.61 | 2.49 | -3.12 | 3.48 | 5.80 | 4.40 | -0.42 |
| ta | 3.90 | 3.40 | 0.00 | -0.42 | 1.56 | 2.76 | 3.08 | 3.24 | 2.92 | -0.37 |
| te | 3.01 | 2.80 | 1.57 | 0.00 | 1.48 | 1.85 | 2.56 | 2.71 | 2.72 | -0.30 |
| ml | 3.90 | 2.40 | 2.06 | -0.06 | 0.00 | 3.69 | 3.00 | 3.14 | 3.19 | -0.31 |
| ur | 6.00 | 2.80 | 1.81 | 0.45 | 2.27 | 0.00 | 3.92 | 4.70 | 4.30 | -0.41 |
| bn | 4.10 | 2.40 | 1.88 | 0.05 | 2.06 | 3.60 | 0.00 | 3.58 | 3.14 | -0.33 |
| gu | 5.20 | 2.80 | 2.54 | 0.17 | 2.47 | 3.78 | 3.88 | 0.00 | 4.07 | -0.50 |
| mr | 4.60 | 3.30 | 2.23 | 0.42 | 1.76 | 2.76 | 3.94 | 3.50 | 0.00 | -0.33 |
| or | 8.48 | 6.09 | 2.07 | 0.13 | 2.33 | -0.66 | 3.95 | 4.68 | 2.47 | 0.00 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| en | 0.00 | 1.40 | -0.42 | 0.39 | -0.12 | 1.30 | 0.60 | 1.30 | 1.20 | 2.20 |
| hi | 0.70 | 0.00 | 0.53 | -0.10 | 0.37 | 0.40 | 0.30 | 0.20 | 1.30 | 1.80 |
| ta | 0.20 | 0.40 | 0.00 | 0.52 | 0.34 | -0.20 | 0.57 | 1.40 | 0.96 | 1.50 |
| te | 1.30 | 1.70 | -0.03 | 0.00 | 0.55 | 0.30 | 1.40 | 2.40 | 1.70 | 1.50 |
| ml | 0.10 | 1.00 | -0.17 | 0.02 | 0.00 | -0.30 | -0.10 | 0.70 | 0.50 | 0.60 |
| ur | 0.70 | 0.40 | 1.33 | 0.70 | 0.06 | 0.00 | 0.20 | 1.40 | 0.80 | 1.50 |
| bn | 0.90 | 0.50 | 0.22 | -0.20 | -0.16 | 0.80 | 0.00 | 0.70 | 1.00 | 1.80 |
| gu | 1.00 | 0.00 | 0.20 | 0.20 | -0.12 | -0.40 | 0.60 | 0.00 | 0.40 | 1.50 |
| mr | 1.10 | 1.10 | 0.53 | 0.10 | -0.25 | 0.30 | 0.40 | 1.30 | 0.00 | 1.70 |
| pa | -0.10 | 0.30 | 0.04 | 0.30 | -0.02 | 0.20 | 0.50 | 1.00 | 0.50 | 0.00 |
| | en | hi | ta | te | ml | ur | bn | gu | mr | pa |

ILCI

**Table 5.11** Improvements (differences in BLEU) in directions by augmenting with PIB vs un-augmented with all remaining data. Reds indicate drop in BLEU scores and blues indicate improvements. We see improvements overall by using PIB as augmented data for training.

### 5.1.6 Release Information

Through this work we will be releasing a multilingual parallel corpora that is obtained from online publicly available documents. The release comprises 407K sentences across 11 languages for the PIB corpus for training and 41K sentences across 10 languages for *Mann Ki Baat* corpus for test. This is the first release of the corpus. This will be made available through the project online website on acceptance of the work. As these online resources would be increasing particularly in low resource Indian languages such as Kannada and Punjabi, this work in corpus construction can be further extended for future releases.


### Acknowledgements

We gratefully acknowledge the online corpora provided by ILCI, WAT-ILMPC, the online publicly available data that we have used in this work and the online open source tools that have facilitated this work.


## 5.2 Automatic Face to Face Translation

Converting video content for consumption for audiences who speak a different language is a space of piqued interest today. This has many applications from naturalizing otherwise unpleasant dubbed movies into other languages, providing access to internationally acclaimed content for courses in English to the non-native speakers who understand their mother-tongue better. In this section, we demonstrate the application of these NMT systems as part of a larger face-to-face translation pipeline involving the text, speech and visual modalities. Figure 5.1 illustrates the pipeline involved starting from video and speech in language $L_A$ to video and lip synchronized speech in language $L_B$.

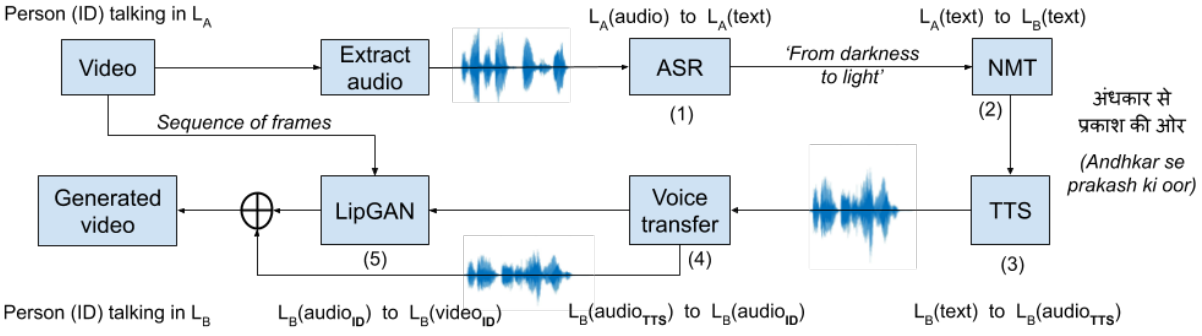


**Figure 5.1** The Face-to-Face translation pipeline. NMT is a critical component where the language-crossing happens.

Concisely, the process can be described as follows: From a video segment, the frames as images and the speech as audio are separated. The audio is then converted to text by an Automatic Speech Recognition (ASR) module, followed by translating it to the required target language $L_B$ using our

NMT system. A Text-to-Speech (TTS) system converts this text to speech. A conditional generative adversarial network (c-GAN), called LipGAN is then trained to synthesis the audio to required lip images which during inference are overlayed on the frame to provide lip-synchronized speech of in the target language. The non LipGAN components are trained independently and requires data for supervised learning. The LipGAN is learned in a self supervised manner from the audio and face frames present aligned in normal video captures. The reader is advised to look at [44] for more details and the non-NMT components of this pipeline. We discuss the the model involved in this work and its standing compared to one other major provider evaluated during the work, demonstrating our model's superior performance below.

### 5.2.1 Translating to target language $L_B$

The transformer architecture [98] enables the state of the art results in several tasks at the time of writing this paper, including machine translation. To bring down vocabulary, while covering the language space minimizing unknowns, we use *SentencePiece*[45][11]. re-implementation of Transformer-Base available in `fairseq-py`[12].

The language pairs we attempt our problem on contains a low resource language, Hindi. To create a nmt system which works well for these along with English, we resort to training a multiway model to maximize learning[65, 2]. We closely follow Johnson et al. [36] in training a multi-way model. The model parameters are shared across all languages and a control token in the input sequence is used to switch language directions while translating. We use only one control token which switches the target language, and let the network infer the source language, as suggested in Johnson et al. [36]. This additionally makes the network robust to code mixed content.

| direction | our-BLEU | Online-G |
|---|---|---|
| Hindi to English | 22.62 | 19.58 |
| English to Hindi | 20.17 | 17.87 |

**Table 5.12** NMT Evaluation Scores.

The multiway model is trained across 7 languages - Hindi, English, Telugu, Malayalam, Tamil, Telugu, Urdu and used in our pipeline to provide translations. In Table 5.12, we report evaluation metrics for language directions which are within the scope of this paper. We indicate the size of training data used and the evaluated scores using Bilingual Evaluation Under Study (BLEU) obtained on a standard test set in each direction available. BLEU is an automated evaluation metric widely used in evaluating machine translation results. The numbers are reported on the test split of IIT-Bombay Hindi-English Parallel Corpus [52]. The test data consist of newscrawled articles, which aren't domain specific. We

[11] https://github.com/google/sentencepiece

[12] https://github.com/pytorch/fairseq

compare against Google Translate in this test set, which is indicated in Table 5.12 as Online-G. We achieve an increase of 3 BLEU points on the test set compared to Google Translate[13]. Translations provided by which are acceptable to concerned demographic at the time of writing this paper.

[13] At the time of the work. Google Translate has improved since, and the values might not reflect the ones reported here.

# Chapter *6*

# Revisiting Low Resource Status of Indian Languages in Machine Translation

Indian language machine translation performance still stay hampered due to the lack of large scale multi-lingual sentence aligned corpora and robust benchmarks. Having established the efficacy of modern day neural and multilingual approaches to machine translation for Indian langauges, in a next step we use these models to enhance both the datasets and the model in an iterative pipeline.

In this chapter, we provide and analyse an automated framework to obtain such a corpus for Indian language neural machine translation (NMT) systems. Our pipeline consists of a baseline NMT system, a retrieval module, and an alignment module that is used to work with publicly available websites such as press releases by the government. The main contribution towards this effort is to obtain an incremental method that uses the above pipeline to iteratively improve the size of the corpus as well as improve each of the components of our system.

Through our work, we also evaluate the design choices such as the choice of pivoting language and the effect of iterative incremental increase in corpus size. Our work in addition to providing an automated framework also results in generating a relatively larger corpus as compared to existing corpora that are available for Indian languages. This corpus helps us obtain substantially improved results on the publicly available WAT evaluation benchmark and other standard evaluation benchmarks.

## 6.1 Introduction

Advances in machine translation, language-modelling, and other natural language-processing has led to a steep increase performance on tasks for many high-resource languages [69, 24]. One major driving factor is many western languages which become test-beds for the methods are already high-resource, which works in favour of methods which are data hungry [43]. The high resource European counterparts have supporting projects like Europarl [41], Paracrawl [8]. These have enabled large scale sentence aligned corpora to be developed. Similar efforts have not been realized for languages in the Indian subcontinent [37]. Evidently, attempts need to be undertaken to improve this situation. Our work

directly addresses this lacuna in the Indian machine translation setting. Specifically, through this work, we aim to achieve the following objectives:

- Provide a large scale sentence aligned corpus in 11 Indian languages, viz. CVIT-PIB corpus that is the largest multilingual corpus available for Indian languages as can be seen from Table 6.1.
- Demonstrate that such a corpus can be obtained automatically with no human effort using iterative alignment method.
- Provide robust standardized evaluation methodology and strong baselines that can be adopted and improved upon to ensure systematic progress in machine translation in Indian languages.

We briefly examine the alternatives to our approach and argue the need for adopting the proposed approach.

**Working at Scale** There have been impressive works for low-resource languages at scale [2, 56, 84], for instance working with 1620 language pairs [84] . However, not all these advances are feasible with regard to the compute resources available to standard academic research groups. Specifically, large models that converge faster, transfer more, and improve performance even for low-resource languages [2, 56] are not trainable on hardware available to many research groups. Hence, we argue that this approach is not viable for Indian machine translation research, at this moment.

**Presently available corpora and baselines** Unfortunately, research in Indian language machine translation suffers from the lack of suitable publicly available models and baselines. Those that are available are rather limited in scope or evaluation. For instance, a widely used corpus that is available is the ILCI Corpus [35]. This corpus has 50K sentences aligned across many languages in the country. However, this corpus is limited to specific domains. The evaluation strategies on this corpora in literature also lacks comparability with no standard test-split. Despite these limitations, the corpus has been used by several reported sources as training data in literature to develop and study Machine Translation [3, 53, 28]. However, this corpus is not useful for applications like KR et al. [44], due to the limitations. The Workshop on Asian Translation (WAT) [64, 62, 61] on the other hand provides a standardized platform for a few languages. Similarly the Workshop on Machine Translation (WMT) [10] from time to time hosts tasks with directions involving Indian Languages. Unfortunately, though several iterations of these tasks have concluded, to the best of our knowledge, there are no trained models that are publicly available at the moment. We summarize and list the presently available corpora in Table 6.1. It is evident that large scale multi-lingual corpora are lacking presently. There are multiple attempts in the past for developing ML solutions for Indian languages, including [12] and other attempts such as [91, 60, 28, 48]. Unfortunately, most of these methods do not evaluate on a standard benchmarks/dataset for reliable comparison of the performance. They also have inferior performance to our approach and do not provide publicly available models. This thereby inhibiting research in the community.

**Proposed approach** We believe, that most methods applicable to the high-resource languages should work just as well in Indian languages in presence of the same amount of data. A simple solution which maintains Occam's razor is to change the low-resource situation, as more content is created online in many Indian Languages which is not pursued as much as it should be. A step taken towards improving the situation in the monolingual corpus space is Kunchukuttan et al. [50].

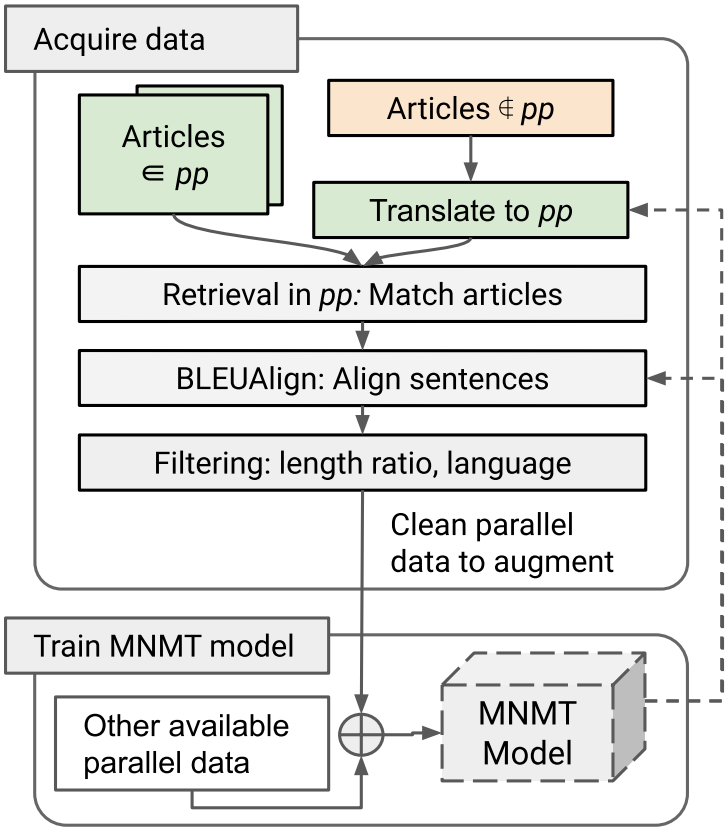


**Figure 6.1** Iterative alignment pipeline used for expanding the corpus for Indian languages. We observe that (i) A better MNMT model leads to better alignment and larger corpus (ii) Larger corpus leads to better MNMT model. We iterate until no further improvement is observed. The dashed lines indicate application of the trained MNMT model. *pp* stands for an arbitrary pivot language.

In this work we demonstrate how, using recent advances in Multilingual Neural Machine Translation (MNMT) [98, 36, 17] in an Expectation Maximization (EM) setup in the face of incomplete data, it is possible to change the status-quo of low resource to produce larger corpus and strong baselines in machine-translation for several Indian languages. A first-step towards this was taken in Siripragada et al. [93] where we presented the CVIT-PIBv0.0 and CVIT Mann Ki Baat corpora. We substantially extend and refine this work through an iterative pipeline illustrated in Figure 6.1. We also co-opt some

ideas proposed in low-resource adaptation for NMT proposed by [90]. In the process, we attain stronger baselines in translating the involved language-directions. Our contributions summarized are as follows:

1. **Low Resource→High Resource**: We extensively study the iterative-alignment methods provided by Sennrich and Volk [89] in the context of CVIT-PIBv0.0 dataset created for Siripragada et al. [93]. Through successful execution of these methods, we increase the corpora size aggregated over all language-pairs from our previous 613K[1] to 1.17M (~ 93% increase) that we term the CVIT-PIB corpus v0.2.

2. **Comparable and Strong Baselines**: We report consequent stronger baselines for MNMT in Indian languages from the improvement in data, validating the utility of the corpora we provide. The final MNMT model covers 11 languages and 110 language-directions with competitive or state-of-the-art performance in 9 tasks on public leaderboards.

3. **Trained models and code**: We release the source-code, trained models and the datasets[2] to further research in this area and to aide applications that could be enabled by a functional MT. To the best of our knowledge, these are the only trained models available for translation in Indian languages at the time of writing this document.

The rest of this document is organized as follows: In Section 6.2, we describe using traditional methods in MT based alignment to improve the parallel corpus across 11 Indian Languages. In Section 6.3 we report stronger baselines for MNMT in Indian languages across many available public tasks.

## 6.2 Iterative Alignment for PIB

We apply the methods in Sennrich and Volk [89] to iteratively improve parallel data. The procedure is analogous to expectation-maximization (EM) algorithm. In the expectation step, we use noisy alignments of parallel sentences from news articles to get a meaningful signal to obtain a better Maximum Likelihood Estimate (MLE) function for the MT model. In the maximization-step the improved MT model is used to obtain stronger alignments. Unlike Sennrich and Volk [89], we use an MNMT model in place of the Statistical Machine Translation (SMT) model. In this section, we provide details about the corpus, the methods used to obtain the same and analysis of its characteristics.

### 6.2.1 Data Sources

To obtain an initial Multilingual NMT (MNMT) system, we rely on the datasets compiled from several sources listed in Table 6.1. We use backtranslation [86] to improve data in Hindi and Telugu.

[1] In Siripragada et al. [93], we report this as 408K considering only English alignments, in this work we consider all-directions.

[2] http://preon.iiit.ac.in/~jerin/bhasha/

The Press Information Bureau (PIB) is used in this work as a source for articles published in several Indian Languages to extract a multiparallel corpus. The PIB is very similar to a newspaper publishing in several languages except with strong one-to-one matches between documents and monotonic sentences which provide more parallel sentences through automatic sentence alignment algorithms. This work uses the same crawled content as Siripragada et al. [93] and focuses on improving the quality of alignments, in an attempt to consequently improve corpus size and performance of the MNMT model.

| Source | #pairs | #lang | type |
|---|---|---|---|
| IITB-en-hi [51] | 1.5M | 2 | en-hi |
| UFAL EnTam [80] | 170K | 2 | en-ta |
| WAT-ILMPC [64] | 800K | 7 | xx-en |
| ILCI [35] | 550K | 11 | xx-yy |
| OdiEnCorp [71] | 27K | 2 | en-or |
| Backtranslated-Hindi | 2.5M | 2 | en-hi |
| Backtranslated-Telugu | 500K | 2 | en-te |
| CVIT Mann Ki Baat[93] | 41K | 10 | xx-yy |
| PMIndia-Corpus[32] | 728K | 13 | xx-yy |
| CVIT-PIBv0.0[93] | 613K | 11 | xx-yy |
| **CVIT-PIBv0.2** | **1.17M** | **11** | xx-yy |

**Table 6.1** Publicly available corpuses for Indian languages. The last group of rows were not used for training. CVIT Mann Ki Baat is used for evaluation purposes only and has overlap with PMIndia Corpus. All other sources are used for training the multilingual model. xx-yy indicates parallel sentences aligned across multiple languages. Last row is the proposed corpus.

We are aware of the existence of PMIndia [32], a source of similar nature and motivation as PIB, but make a conscious choice not to use it in this work to prevent data-leakage issues of possible overlap with one of our test-sets CVIT Mann Ki Baat [93].

### 6.2.2 Iterative Alignment

Our iterative alignment procedure requires document and sentence alignment algorithms, an MNMT formulation which is trained again with refreshed data in each iteration. We describe the constituent components illustrated in Figure 6.1 and describe the iterative procedure ahead.

#### 6.2.2.1 Text Processing

**Text Cleaning and Standardization** We allow for noise on the web in the pipeline and avoid any linguistic features. There is noise present in documents generated in the past with Indian languages content. A desideratum for retrieval and matching is that the model is capable of handling such noise. This could be unicode issues, non-standard or normalized text which is present all-across sources on the web. Some amount of noise is mitigated by past works [12] by standardising unicode, scripts etc[3].

**Tokenization** We use SentencePiece [47] which can be used as a tokenizer which tokenizes into subword-units. The subwords which cover a corpus are decided optimizing likelihood of a unigram language-model over a large corpus and candidate subwords in EM steps. Recent works [24, 66] addressing high-resource western-languages follow a joint vocabulary of some 32K-64K subwords as a subword model creation hyperparameter. We observed in our early experiments that in the presence of huge imbalance of data among languages and the huge difference in scripts unlike major European languages, for example, this approach leads to subwords which reduce to characters for the less represented languages. In order to avoid the artifacts from such a subword learning strategy, we instead choose 4K subwords for each of the languages involved and take a union of these to generate the final vocabulary[4]. The process results in a vocabulary of 40K subword-units tokens for 11 languages which we maintain fixed across all iterations. We note that the artifacts can also be mitigated by a temperature based sampling for sentences among languages as Aharoni et al. [2].

**Filtering parallel pairs** To obtain a filtered corpora at every iteration, we allow only sentence-pairs into to the training pipeline where source length to target length ratio is in $[0.5, 2.0]$. We also use `langid`[5], a language identifier through writing script to filter sentences with foreign language tokens.

#### 6.2.2.2 Alignment Algorithms

To perform document alignment we translate the articles to a common pivot language. We use the translations to rank candidate-matches in the pivot language. SentencePiece tokenization of each sentence eliminates requirement of a curated stop-words list, enabling us to compute similarity in the search space of any desired pivot language. Cosine similarity on the *term frequency - inverse document frequency* (tf-idf) [13] is put to use to rank retrieved articles in the space of pivot language. Search space to find and rank candidate articles matching a translation is restricted to only in a vicinity of dates (2 days) of posted news articles.

Upon obtaining aligned document pairs, we use Bleualign [88], an MT based sentence alignment algorithm. Other conventional sentence length based alignment algorithms such as Gale-Church [25]

[3] https://github.com/anoopkunchukuttan/indic_nlp_library

[4] This design choice is partially inspired by the reasoning and supporting experiments in Sennrich and Zhang [90].

[5] `https://github.com/saffsd/langid.py`

also exist, but we rely on MT based alignment as the performance of the NMT model increases with every iteration resulting in better sentence alignment. Bleualign also aggressively filters reducing false matches [8].

| | Model | Languages | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | hi | ta | ml | mr | gu | te | or | bn | pa |
| Retrieval Accuracy | M2M-0 | 76.77 | 54.12 | 45.63 | 34.05 | 52.52 | 24.06 | - | - | - |
| | M2EN-1 | 86.91 | 71.36 | 69.77 | 47.57 | 63.67 | 62.3 | - | - | - |
| | M2EN-2 | 92.39 | 80.84 | 80.4 | 51.89 | 75.86 | 70.05 | - | - | - |
| *xx*→*en* BLEU | M2M-0 | 17.6 | 9.5 | 11.1 | 11.4 | 14.5 | 7.1 | 3.1 | 8.9 | - |
| | M2EN-1 | 21.4 | 14.4 | 15.6 | 16.8 | 20.1 | 9.4 | 6.4 | 15.1 | - |
| | M2EN-2 | 22.4 | 15.3 | 16.8 | 17.7 | 21.4 | 9.9 | 17.1 | 16.4 | - |
| | M2EN-3 | 22.0 | 15.6 | 16.9 | 18.2 | 21.4 | 10.2 | 19.5 | 16.6 | - |
| Corpus Size | M2M-0 | 156.3K | 61.0K | 17.0K | 40.0K | 25.5K | 6.0K | 9.1K | 21.6K | 26.3K |
| | M2EN-1 | 189.2K | 73.7K | 28.7K | 71.6K | 26.3K | 5.3K | 20.3K | 42.4K | 24.6K |
| | M2EN-2 | 195.2K | 87.1K | 32K | 81.0K | 29.4K | 5.7K | 58.5K | 48.3K | 27.1K |
| | Increment | 38.9K | 26.1K | 15.0K | 40.8K | 3.9K | 0 | 49.4K | 26.8K | 0.8K |

**Table 6.2** Incremental improvements in Accuracy, *xx*→*en* BLEU scores on *CVIT Mann Ki Baat* and Corpus size. We observe increments in retrieval accuracies consistent with increase in BLEU scores. We increment the initial version of the PIB corpus with an additional *201K* sentences aligned to English. (*CVIT Mann Ki Baat* does not contain Punjabi.)

#### 6.2.2.3 Multilingual Neural Machine Translation (MNMT) Model

We use fairseq [67] for training a Transformer-Base [98] based MNMT system. The model we begin with is same as our first multilingual model in Siripragada et al. [93]. However, unlike Siripragada et al. [93], to refine the CVIT-PIBv0.0 dataset further, we choose a many-to-English model formulation trained to translate only from non-English languages to English. This is advantageous because (1) it enables faster training and retraining time, (2) the setting provides more capacity to English decoding which improves translation performance and consequently – retrieval in English. The model parameters - encoder, decoder and embeddings are shared among all languages. We additionally use tied embeddings [79] at the encoder and decoder. In this work, we denote our first many to many model with no CVIT-PIBv0.0 dataset augmentation as M2M-0, the following many to English models with incremental dataset variations as M2EN-1, M2EN-2, M2EN-3. We additionally consider the best model from Siripragada

et al. [93] after training with CVIT-PIBv0.0 dataset augmentation, denoted in this work as M2M-1, to attempt another translation to a non English and possibly related-language as an alternative for retrieval. In the next section, we also consider M2M-3, a model capable of translating all directions with the CVIT-PIBv0.2 dataset after iterative refinement.

#### 6.2.2.4 Iterations

In each iteration, we initialize training with the model from previous iteration (warm-start). This helps to reduce training time when compared to a model training from scratch (cold-start) as we benefit from learning in the previous iterations. To maintain a constant increment in articles, we set a threshold on keeping a retrieved candidate at a constant value for each language. We observe that the scores improve with successive iterations, consequently obtaining more matching documents. We stop the iteration process at a stage of diminishing returns, i.e there is no prospect of a justifiable increase in corpus size (See Figure 6.2). Note that in future, further expansion through an increasing number of documents is always possible.

### 6.2.3 Discussions

The many-languages involved and the disparity in sizes in training lead to a setting where we can dissect and study several aspects. First we study the iterative alignment process, comparing retrieval accuracy, BLEU scores and increase in corpora side by side.

We track BLEU [70] for the MT model, a *pseudo* retrieval accuracy for the retrieval pipeline and count of successful sentence alignments for increase in corpora over iterations. We employ the following technique to arrive at our *pseudo* retrieval accuracy. The articles which match in dates and ministry information from meta-information are collected. In many languages, there are enough true-positive matches to report the consequent evaluations as a proxy to retrieval accuracy.

These numbers reported in Table 6.2 indicate mostly consistent trends reflecting improvement in BLEU scores in translating to English for the CVIT Mann Ki Baat test set, retrieval accuracies and consequently the resulting acquired data-sizes. The BLEU scores improve with the addition of more data, while the retrieval accuracy and data sizes improve with updating the parallel-corpus generated from PIB using the higher-performing MNMT model. Through successive iterations, the corpora increases in size and gets refined. We observe in Table 6.2, most increase for the languages with already good data (Hindi). In two iterations, we add a net 201K sentences aligned to English on top of the 408K (49% increase) sentences in the previous release, CVIT-PIBv0.0. But however, it is worth noting that some languages which are aided by the transfer and the new data have almost increased an order in sizes in iterations later (Marathi, Oriya and Bengali). We describe the details of the new parallel corpora in Table 6.3. Once the parallel corpus extraction from PIB is saturated, we obtain the multiparallel corpus by getting sentence alignments amongst other languages by bridging through the English part of the existing English centric

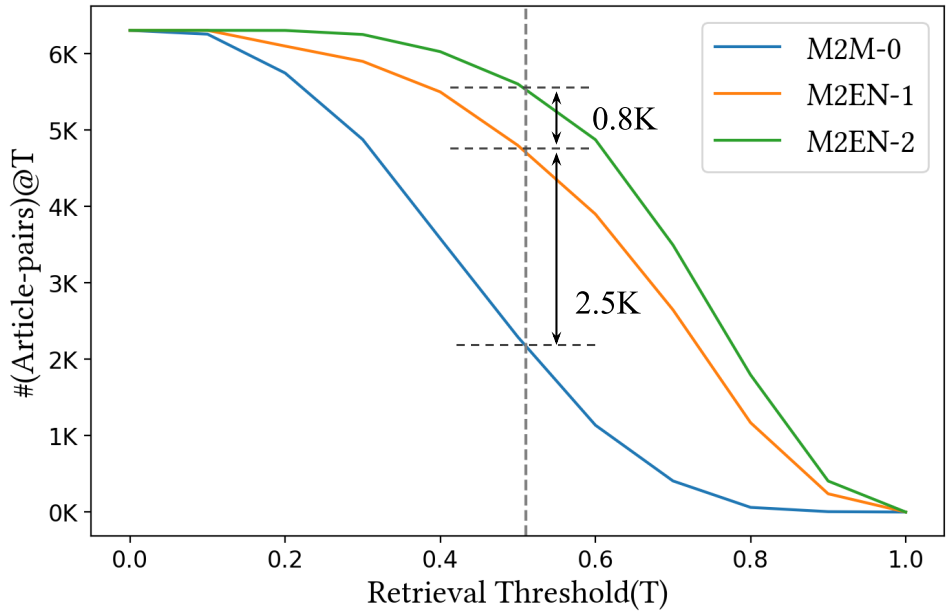


**Figure 6.2** Figure illustrating number of articles on the Y-axis obtained at a given threshold on X-axis. We maintain a constant threshold (dotted line) of *0.51* across model iterations M2M-0, M2EN-1 and M2EN-2. In case of Marathi, when compared to M2M-0 we acquire an additional 2.5K article pairs using M2EN-1 and 0.8K more using M2EN-2. From the graph, we observe saturation in articles pairs after iteration 2 indicating a point of diminishing returns.

data. The process ends ups providing an additional 564K (93% increase) sentences to the previous 609K sentences aligned across languages, resulting in a corpora size of *1.17M*.

In the case of Oriya, we observe no date-based matches. This leads to depending entirely on accurate document pair retrieval for extracting a corpus of reasonable alignment accuracy. The preliminary efforts used to retrieve and align data in Siripragada et al. [93] comprises of Bible [71] which has been observed to transfer poorly to other domains consequently to poor sentence level alignment accuracies for Oriya in our earlier models Siripragada et al. [93]. However, with every iteration in Tables 6.2 and 6.4 we observe increase in BLEU scores when translating to English. This leads to better retrieval and sentence level accuracies.

The languages Marathi (mr), Gujarati (gu), Punjabi (pa) are similar to Hindi and exhibit high BLEU scores (Table 6.6) when translated to Hindi. They are also known to be similar[49], so we experiment with hi as a pivot language. However, we found poor retrieval performance when compared to pivoting through English. We observed that Hindi articles are much more elaborate in content while the mr, gu, pa versions are often summarized. This is evident when considering examples of Gujarati articles, as PIB offices of Gujarat are responsible for posting the articles in Gujarati and their respective English translations. We illustrate this phenomenon through an example in Figure 6.3, where we observe higher

retrieval scores overall when compared to Hindi-based retrieval. The above analysis points to the success and a potential use-case of our model in being able to deliver consistent content across all languages for websites like PIB.

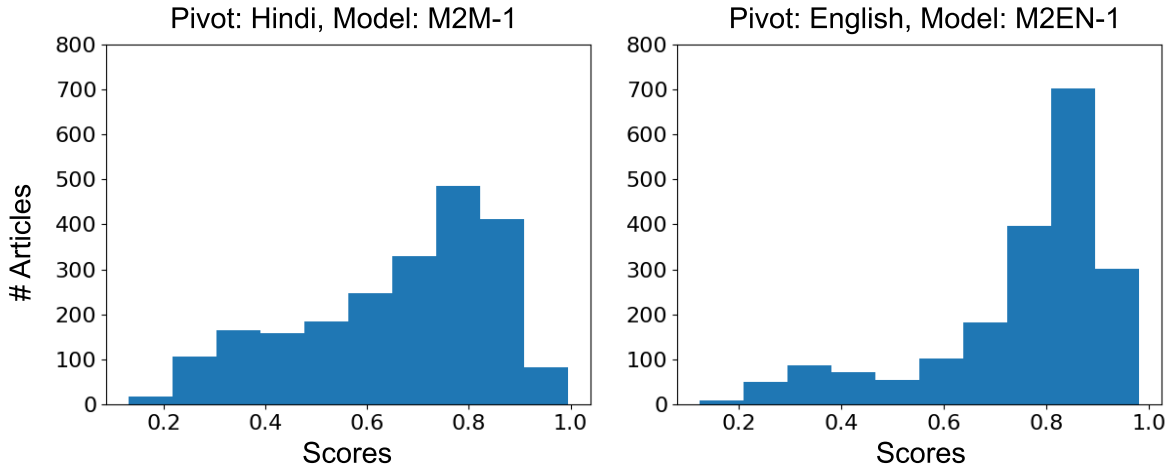


**Figure 6.3** Retrieval scores of *Gujarati*. Left bar chart indicates retrieval scores in case of model M2M-1 and pivot language *hi*. Right chart indicates scores in case of M2EN-1 and pivot *en*.

In our experiments, we had left Urdu out from the iterative alignment procedures due to unresolved bugs in the text processing in the pipeline. This lead to no improvement in Urdu to English corpus sizes. However, this paves the way for a case study to evaluate the effect of the improvements in other multilingual language-pairs in translating to and from Urdu as we include Urdu in the performance evaluations. We notice that as the remaining resources improve and better alignments are put in place, and with no further Urdu data enhancements, there are improvements in the BLEU scores of language pairs involving Urdu, observed in Table 6.4.

To summarize, we provide a new corpus and a method to expand the corpus from publicly available sources.

- We make a new large sentence aligned corpus (CVIT-PIBv0.2) available for researchers, as a result of the iterative alignment described above. The detailed analysis of the iterative alignment procedure is provided in Table 2.

- This new corpus is possibly the largest sentence aligned multi-lingual corpus for Indian languages as can be observed through Table 1. [93]) in number.

| | en | hi | ta | te | ml | ur | bn | gu | mr | or | pa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| en | | 195208 | 87113 | 5752 | 31974 | 45344 | 48354 | 29421 | 80760 | 58461 | 27117 |
| hi | | | 44031 | 3083 | 17819 | 11695 | 24849 | 19730 | 45950 | 36317 | 11442 |
| ta | | | | 3218 | 15029 | 4964 | 19175 | 16934 | 33636 | 27668 | 9150 |
| te | | | | | 2543 | 415 | 1883 | 2625 | 2627 | 1834 | 1220 |
| ml | | | | | | 2378 | 9940 | 10132 | 14474 | 9843 | 4961 |
| ur | | | | | | | 3795 | 2397 | 4941 | 3209 | 5584 |
| bn | | | | | | | | 10554 | 19914 | 14606 | 5332 |
| gu | | | | | | | | | 17169 | 13682 | 5581 |
| mr | | | | | | | | | | 31377 | 9601 |
| or | | | | | | | | | | | 6813 |

**Table 6.3** Multilingual shared content across language pairs for CVIT-PIBv0.2. Rows and columns indicate language pairs. The highlights are proportional to the change after the iterative alignment process, reds indicating decrease and blues indicating increases in corpora sizes compared to the previous release v0.0. Evident from the table, we notice major increments in Marathi, Oriya and other languages. Tamil and Hindi which we had enough to be considered mid-to-high resource gain significant number as well. The maximum decrease is -283 for Telugu, which is negligible compared to the improvements of the order of ten-thousands in many language-pairs.

- Table 3 shows the corresponding increase in size and percentage increase from our previous effort using the same document set as in [93] with the increment obtained only through the proposed procedure. We further provide strong baselines that are presented in the next section.

- We also hint at a method that can allow continuous expansion of this corpus and eventually enabling a class of recent language processing methods on Indian languages.

Having described the process of iterative refining and enriching parallel-corpus resources for Indian Languages, we use the resulting refined corpora in training two models for obtaining baselines - one many-to-many (M2M) and the other many-to-English (M2EN). Since these are the next iterations, we label these M2M-3 and M2EN-3. The two models are used to establish strong baselines for the 11 languages and consequent 110 directions involved which we describe in the next section.

## 6.3 Stronger and Repeatable Baselines

It is important to further research to first take stock of where we are, and often simple baselines which compete in comparison to sophisticated methods serve this exact purpose [100, 4]. We consider the possibility using standard approaches which work for high-resource languages to provide such baselines for translation among Indian languages. Our previous work Philip et al. [74] reports baselines for multilingual translation models for 5 Indian Languages and English. Further improvements are brought about in Siripragada et al. [93] in several tasks adding more languages. It is also important to take care that these are "simple baselines" which are comparable [77], leaving plenty of room for improvements ahead.

Addressing simplicity, through the aforementioned works and this one, the model architecture and hence the learner's "capacity" is kept constant (Transformer-Base). Without increasing capacity, the experiments continue taking on all available translation tasks in Indian languages. A task here corresponds to language-directions or domains of datasets. The model we use is expected to be as general purpose as possible, after improving the data situation. There are no linguistics based priors in our methods or explicit handling of noise. This provides room for linguistics based improvements to build on simultaneously raising a call to revisit some older propositions. Addressing comparability, we comprehensively cover and compare with test-sets in prior-art and the shared tasks.

### 6.3.1 On Repeatability of Objective Evaluations

We address two important aspects (1) standard test-sets available and (2) reproducible evaluations.

#### 6.3.1.1 Test Sets

We identify two class of test-sets among Indian languages, (1) which corresponds to ILCI which many early works evaluated translation quality on and (2) associated with the WAT or WMT tasks, which provide a leaderboard and standardized evaluations for comparison. To cover limitations of these test-sets, we propose a new test-set CVIT Mann Ki Baat in Siripragada et al. [93]. We proceed to summarize how we compare to past work reporting numbers on these test-sets.

#### 6.3.1.2 Comparable reports of BLEU Scores

Post [76] addresses several issues of reproducibility and fair comparisons in reporting BLEU scores. In this work, we make our evaluations consistent with WAT leaderboard and provide a package to reproduce the procedure locally[6]. For WMT tasks, we report the values from the portal[7](gu-en) and the

[6] https://github.com/jerinphilip/wateval

[7] http://matrix.statmt.org/matrix

| direction | Model | IITB | UFAL | OEC | WAT-ILMPC | | | | | | W19 | W20 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | hi | ta | or | hi | ta | te | ml | ur | bn | gu | ta[1] | ta [2] |
| *en→xx* | M2M-0 | 19.83 | 6.78 | 4.29 | 19.9 | 10.8 | 16.1 | 7.2 | 13.2 | 9.7 | 6.4 | 3.5 | 4.8 |
| | M2M-1 | 20.52 | 7.31 | 5.26 | 20.3 | 11.6 | 16.6 | 7.9 | 16.5 | 9.6 | 9.5 | 4.2 | 6.0 |
| | M2M-3 | 21.20 | 7.22 | 4.78 | 20.9 | 11.9 | 17.1 | 7.7 | 15.8 | 10.1 | 11.3 | 4.9 | 7.1 |
| *xx→en* | M2M-0 | 21.94 | 18.64 | 11.0 | 27.9 | 17.8 | 21.6 | 11.9 | 20.5 | 16.6 | 17.9 | 12.9 | 12.7 |
| | M2M-1 | 22.48 | 19.76 | 10.84 | 28.3 | 18.6 | 22.5 | 12.7 | 21.1 | 16.7 | 23.6 | 14.3 | 13.9 |
| | M2M-3 | 23.07 | 19.87 | 12.07 | 29.0 | 19.2 | 24.0 | 12.8 | 21.1 | 17.4 | 25.6 | 15.9 | 15.3 |
| | M2EN-1 | 23.83 | 23.38 | 13.07 | 31.3 | 21.2 | 25.7 | 14.2 | 23.4 | 18.8 | 22.8 | 15.5 | 15.0 |
| | M2EN-2 | 24.65 | 25.32 | 15.62 | 32.9 | 23.2 | 28.1 | 15.7 | 24.5 | 20.0 | 24.5 | 16.6 | 16.3 |
| | M2EN-3 | 25.26 | 26.08 | 17.76 | 34.1 | 23.8 | 29.5 | 16.4 | 25.9 | 20.6 | 25.3 | 16.7 | 16.4 |

**Table 6.4** We report BLEU scores on available publicly available benchmark tasks for Indian Languages. The results on these benchmarks often have models that are specially tuned for various language pairs. We do observe that we obtain state of the art results on 3 of the language pairs and are competitive to other works that are more specific in most cases. This is despite not being specially tuned for these settings. OEC stands for OdiEnCorp, W19, W20 stands for WMT19 and WMT20. [1] stands for test-split and [2] stands for dev-split respectively.

SacreBLEU[8] signature (ta-en). We make the hypotheses generated among all test-sets available[9] in case of a requirement to re-evaluate and compare with non-BLEU metric.

### 6.3.2 Results and Discussions

We begin by discussing the general merits of the model especially coming out of the multilingual formulations, and proceeding to elaborate on where we stand with regard to existing literature. For translating in to English, our best M2EN model (M2EN-3) provide a cumulative improvement of +21 BLEU and an average improvement of +2.6 BLEU, compared to the previous best known multilingual model M2EN-1 [93] with a similar coverage of languages, with the same model capacities on CVIT Mann Ki Baat test set. This clearly points to the improvement consequent of change in data situation in the involved Indian languages. In Table 6.6, we report BLEU scores of the M2M-3 in a grid indicating the performance in the language-directions the model applies. We also highlight the improvement magnitude in color. An average improvement of +4 BLEU points in translating to Urdu and +2 when translating

[8] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.12

[9] https://github.com/shashanksiripragada/generation-results

from Urdu can be observed in 6.6, despite the minimal change in parallel-sentences improvement in Urdu, which is gain solely from multilingual formulations.

The M2EN model being stronger in the tasks it is trained towards compared to the corresponding M2M model, and is consistent with what is established in literature[2]. A reasoning for this disparity is not enabling temperature based sampling as in Aharoni et al. [2] to balance out all language-pairs and correct for the imbalance in huge number of English aligned sentences existing in the training data. In Bapna and Firat [9], corrections for the imbalance are observed to have led to degradation in high-resource languages. Note that both models are trained with the same capacity, and gain in BLEU scores and capability in translating more languages is a major advantage of the M2M model.

To study generalization, we take the models from the iterative procedure described in Section 6.2 and evaluate their performance on all available test-sets. The results can be observed in Tables 6.2 and 6.4. From Table 6.4, it is evident that the augmentation with CVIT-PIB data does not degrade the model in another task, but only improves.

Over the span of a few incremental works, we have significantly improved Hindi to English translation by a margin of +3 BLEU since Philip et al. [74], obtaining higher numbers by using simple methods. This is unlike many rounds of distillation, hyperparameter optimization done by other groups to reach similar range of values[64]. The other languages also have similar improvements in all directions.

Despite being not trained with the provided Gujarati data and using the data from ILCI and CVIT-PIBv0.2 corpus, we are able to achieve a BLEU score of 25.3 and 25.6 with our models M2EN-3 and M2M-3 respectively, competitive to the best BLEU score of 26.9 [57] in WMT 2019 gu-en task. The best performing model did several rounds of backtranslation, distillation and multilingual formulation leveraging Hindi. Among these, we have only taken advantage of Multilingual formulation and on top, pure data augmentation at the moment. A caveat here is that we have not put in efforts into filtering the test-data from the training data. But our corpus collection is independent of the news-sources WMT19 used and the Gujarati-English directions are as good as the claim, also supported by the results on CVIT Mann Ki Baat and ILCI test sets.

The BLEU scores on CVIT Mann Ki Baat test-set are provided in Table 6.6. We notice the most improvements in Oriya involved directions. Our overall multilingual model seems to have improved at M2M-3 in comparison to M2M-1. Despite not enforcing any linguistic priors, we get strong performance in many related languages.

### 6.3.3 Comparison with Previous Works

**Non-standard comparisons** Comparison is not standardized since previous methods evaluate on their own test set or on non-standard splits of the ILCI corpus that are not publicly available. This could be common in the initial stages of research for any community. Our work addresses it by evaluating it on more standard benchmarks with clear publicly available test splits. Kunchukuttan et al. [53] attempts to build a collection of SMT models covering 11 Indian languages, similar to this work, except training and testing on splits from ILCI corpus (2K test-sentences, 500 for validation and remaining for training).

| Work | IITB-hi-en | | WAT-ILMPC | |
|---|---|---|---|---|
| | *en→hi* | *hi→en* | *en→hi* | *hi→en* |
| SMT [51] | 11.75 | 14.49 | - | - |
| NMT [51] | 12.23 | 12.83 | - | - |
| Saini and Sahula [82] | 18.22 | - | - | - |
| Philip et al. [73] | **21.57** | 20.63 | - | - |
| Dabre et al. [18] | | | **29.65** | 31.51 |
| Goyal and Sharma [29] | - | 18.64 | - | - |
| Philip et al. [74] | 20.17 | 22.62 | 26.25 | 31.55 |
| Siripragada et al. [93] | 20.52 | 22.48 | 20.3 | 28.3 |
| Proposed Methods | 21.20 | **25.26** | 20.92 | **34.09** |

**Table 6.5** Comparison with publicly available baselines for English to Hindi and vice versa.

However, the split is not available. Goyal et al. [28] once again report numbers on ILCI, using a similar split strategy as Kunchukuttan et al. [53]. Murthy et al. [60] compares with a similar test-set of 2K ILCI sentences. Similarly, [28] uses the ILCI test-set with a similar strategy to apply a formulation taking advantage of "related-languages". In our experiments we have found ILCI to be domain-specific (health, tourism) and providing false perception of high-scores by models which fail to generalize [93]. Due to the lack of reproducibility and comparability and the known demerits, we do not recommend future comparisons on any arbitrary ILCI test-set for benchmarking general purpose translation systems.

**Comparison on Hindi and English** Hindi is the Indian language that has received highest attention with multiple attempts for translatiion to and from English. The results for comparison for Hindi-English on publicly available standard benchmarks that can be accessed are provided in Table 6.5. We obtain the highest scores for two of the tasks, i.e. IITB and WAT-ILMPC Hindi-English evaluations. Note that the highest BLEU score for IITB English to Hindi was obtained by our previous approach [73].

**Comparison on Public Leader Boards** We provide detailed results for our method on publicly available leaderboards in Table 6.4 and Table 6.6. These can be used for comparisons and evaluations by various methods. As mentioned previously, these are on WAT tasks, WMT Tasks and CVIT Mann Ki Baat evaluation set. In all these tasks our models perform well obtaining state of the art results for several tasks. For instance, we obtain state of the art BLEU score of 17.76 on OdiEnCorp that is much higher than the previous state of the art of 8.6 and 23.8 BLEU score on WAT-ILMPC for Tamil to English that is competitive with the state of the BLEU score of 24.31.

| | bn | en | gu | hi | ml | mr | or | ta | te | ur | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| bn | | 16.27 | 13.75 | 20.35 | 5.18 | 10.19 | 11.97 | 4.44 | 4.29 | 20.65 | 22.96 |
| en | 7.93 | | 11.80 | 16.21 | 5.14 | 9.46 | 8.75 | 4.56 | 4.45 | 18.44 | 14.28 |
| gu | 12.03 | 21.00 | | 42.55 | 6.55 | 16.21 | 14.62 | 5.59 | 5.54 | 39.29 | 29.88 |
| hi | 12.79 | 21.01 | 33.53 | | 7.16 | 17.15 | 15.21 | 5.78 | 6.30 | 45.50 | 32.36 |
| ml | 8.17 | 15.88 | 11.71 | 18.13 | | 9.74 | 10.27 | 4.87 | 4.88 | 18.63 | 19.95 |
| mr | 10.04 | 17.50 | 18.34 | 25.06 | 6.09 | | 11.83 | 4.98 | 4.82 | 25.57 | 25.77 |
| or | 12.88 | 17.53 | 17.29 | 25.28 | 4.77 | 12.42 | | 4.57 | 4.54 | 22.00 | 56.14 |
| ta | 6.78 | 14.59 | 10.06 | 15.78 | 5.16 | 8.39 | 9.10 | | 3.87 | 16.43 | 19.78 |
| te | 7.20 | 10.71 | 10.18 | 15.55 | 5.63 | 8.24 | 8.36 | 4.43 | | 12.00 | 15.11 |
| ur | 10.05 | 21.53 | 25.57 | 44.92 | 6.43 | 13.99 | 10.31 | 4.66 | 5.60 | | 20.84 |
| Δ | 12.65 | 13.42 | 21.57 | 23.93 | 6.77 | 21.97 | 99.51 | 6.00 | 9.67 | 41.58 | |

**Table 6.6** BLEU scores of M2M-3 model on multilingual test set Mann-Ki Baat. Rows correspond to source languages and columns target languages. The colors indicate improvement (blue) or degradation (red) in comparison to M2M-1 [93]. We observe cumulative increment of +257 BLEU across all language pairs and a median increment of +1.25. The cumulative changes in translating to or from a given language in comparison to M2M-1 are provided under Δ header. It can be observed that related languages end up with higher BLEU scores without having to add the prior in the model formulation - e.g (hi, gu), (ur, gu), (ur, hi). Closely behind, there is (mr, hi) ahead of other language pairs. While there seems to be a decrease in *en*→*xx* between M2M-1 and M2M-3 on Table 6.4, the overall improvements here indicate that it is not the general case.

### 6.3.4 Summary: Stronger Baselines and Public Benchmarks

Through this work:

- We establish strong baselines for machine translation for Indian languages.
- Our multilingual model outperforms the previous works and even the carefully handcrafted MT for specific language pairs.
- This work also possibly acts as the first NMT model that is publicly available for research. We hope this will spur more research within the research groups, specially in India.
- Our code, models and data splits are publicly available for reproducibility in the future.

- Our corpus is now getting used in the WMT and WAT (international and Asian premier machine translation forums), demonstrating the utility.

## 6.4 Discussions and Future Directions

In this work, we have made contributions to change the low-resource status of several Indian Languages for Machine Translation. Specifically (i) We introduced a large corpus that can enable Deep Machine Translation and associated research for these languages. (ii) Domain of these sentences allow to cover wider topics and practically more useful. More importantly, we established the utility of an algorithm that can help to grow the size starting from this state. More data will lead to better models, that in turn better alignment and more data.

A challenging question will be the applicability of this method for online resources that are not really created by explicit translation. We believe, that a solution to this problem may not be too far from here. The methods used for search and alignment in this paper can be extended to use newspapers and news specific to a time-window in a weakly supervised setting with minimal human effort to enhance the parallel-corpus further. Using embeddings trained to mine parallel sentences have shown promise for High Resource Languages, which we will incorporate into this pipeline in the future. The meta-information on the stored PIB articles opens up possibilities to study document translation and active learning problems, left for future work.

Our M2EN models have high BLEU scores which allows for an application of backtranslation [86, 24] to improve the numbers further in the opposite direction. Kim et al. [39] reports scenarios where unsupervised NMT methods fail for the low-resource Gujarati-English pair due to limitations, and the enhancement of resources here implores a revisit.

With high-performing NMT systems to English from Indian languages, it is possible to create datasets and corpus for use in downstream tasks and by using the models provided by this work to further the research in Indian Languages.

# Chapter *7*

# Conclusion

In this work, we have demonstrated building a machine translation system with neural and multilingual methods, and its success in both high-resource and low-resource Indian Languages. First, we describe and demonstate the use-case of neural approaches in a single-language pair involving Indian languages in a data-scarce but narrow domain setting. Further, we show success in the general domain English-Hindi language pair to achieve state-of-the art in several Indian languages tasks on public leaderboards. The state-of-the art Hindi-English models have been further used in [44] to enable automatic face-to-face translation. For low-resource languages and language-directions spoken in the subcontinent, we have demonstrated the success of NMT and parameter sharing in bringing about high performances.

Using the above systems, to mitigate the lack of parallel datasets and further progress in this field with regards to Indian languages we have created and released parallel corpora for 10 Indian languages and demonstrated its effectiveness in multilingual machine translation. Our systems are set to collect more sentence pairs as the more articles are published, similar to Europarl [41]. The nature of the corpora opens avenues to explore on how to efficiently use weakly aligned formulations to improve the corpora iteratively. Many languages that are related have better translation performance between them compared to English, examples being Hindi-Punjabi, Hindi-Marathi, Hindi-Urdu. MT based alignments between these can be used to create larger mappings amongst the sentences in both. In addition to the above application, the NMT model demonstrates its utility as a critical component in a Face-to-Face translation pipeline.

Iterative application of such methods can lead to a refined corpus, through minimum human annotation. As we close this thesis, we perform iterative application to extract maximum parallel-sentences from the existing databases and in the process set-up our implentation to provide more parallel pairs as our source publishes in perpetuity. Finally, we comprehensively benchmark the multilingual models on 12 known public tasks across language-directions, to lay strong foundations for future machine translation research in Indian languages to build on.

# Related Publications

- CVIT-MT Systems for WAT-2018 Jerin Philip, Vinay P. Namboodiri and C.V. Jawahar Proceedings of the 5th Workshop on Asian Translation 2019
- A Baseline Neural Machine Translation System for Indian Languages, Jerin Philip and Vinay P. Namboodiri and C.V Jawahar, arXiv preprint arXiv:1907.12437, 2019
- Towards Automatic Face-to-Face Translation, Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha and Vinay P. Namboodiri and CV Jawahar, Proceedings of the 27th ACM International Conference on Multimedia, 1428–1436, 2019, ACM
- CVIT's submissions to WAT-2019 Jerin Philip, Shashank Siripragada, Upendra Kumar, Vinay P. Namboodiri and C.V. Jawahar Proceedings of the 6th Workshop on Asian Translation 131–136 2019
- A Multilingual Corpora collection effort for Indian Languages, Jerin Philip, Shashank Siripragada, Vinay P. Namboodiri and C.V. Jawahar Proceedings of the LREC, 2020
- Revisiting Low Resource Status of Indian Langauges in Machine Translation, Jerin Philip, Shashank Siripragada, Vinay P. Namboodiri and C.V. Jawahar, Under Review, 2020

**Other Publications**

- A Cost Efficient Approach to Correct OCR Errors in Large Document Collections, Das, Deepayan and Jerin Philip and Mathew, Minesh and Jawahar, CV, Proceedings of International Conference on Document Analysis and Recognition (ICDAR), 2019

# Bibliography


[1] Ruchit Agarwal, Mihir Sekhar, and Dipti Mishra Sharma. 2017. Three-phase training to address data sparsity in Neural Machine Translation. *ICON*.

[2] Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

[3] Gary Anthes. 2010. Automated translation of indian languages. *Communications of the ACM*, 53(1):24–26.

[4] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2019. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*.

[5] Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[7] Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):472–482.

[8] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

[9] Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548.

[10] Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

[11] Pushpak Bhattacharyya. 2015. *Machine translation*. CRC Press.

[12] 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*, pages 17–20, San Diego, California. Association for Computational Linguistics.

[13] Christian Buck and Philipp Koehn. 2016. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

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

[15] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

[16] Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv preprint arXiv:1806.04402*.

[17] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *arXiv preprint arXiv:2001.01115*.

[18] Raj Dabre, Anoop Kunchukuttan, Atsushi Fujita, and Eiichiro Sumita. 2018. NICT's participation in WAT 2018: Approaches using multilingualism and recurrently stacked layers. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

[19] Sandipan Dandapat and William Lewis. 2018. Training deployable general domain mt for a low resource language pair: English–bangla.

[20] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language Modeling with Gated Convolutional Networks. In *International Conference on Machine Learning*, pages 933–941.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[22] Omkar Dhariya, Shrikant Malviya, and Uma Shanker Tiwary. 2017. A hybrid approach for hindi-english machine translation. *Information Networking (ICOIN)*.

[23] Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, and Pushpak Bhattacharyya. 2014. The iit bombay hindi-english translation system at wmt 2014. *Workshop on Statistical Machine Translation*.

[24] 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*, pages 489–500.

[25] William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

[26] GV Garje and GK Kharate. 2013. Survey of machine translation systems in india.

[27] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

[28] Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online. Association for Computational Linguistics.

[29] Vikrant Goyal and Dipti Misra Sharma. 2019. LTRC-MT simple & effective Hindi-English neural machine translation systems at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 137–140, Hong Kong, China. Association for Computational Linguistics.

[30] Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural moachine translation. *arXiv preprint arXiv:1906.07286*.

[31] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

[32] Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

[33] Rejwanul Haque, Sudip Kumar Naskar, Josef Van Genabith, and Andy Way. 2009. Experiments on domain adaptation for english-hindi smt. *PACLIC*.

[34] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics.

[35] Girish Nath Jha. 2010. The TDIL Program and the Indian Language Corpora Intitiative (ILCI). In *LREC*.

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

[37] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

[38] Nadeem Jadoon Khan, Waqas Anwar, and Nadir Durrani. 2017. Machine translation approaches and survey for indian languages. *arXiv preprint arXiv:1701.04290*.

[39] Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

[40] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *ACL*.

[41] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. Citeseer.

[42] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. *ACL*.

[43] Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

[44] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1428–1436.

[45] 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)*, volume 1, pages 66–75.

[46] 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)*, pages 66–75. Association for Computational Linguistics.

[47] Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

[48] Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Learning variable length units for smt between related languages via byte pair encoding. *arXiv preprint arXiv:1610.06510*.

[49] Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent. *arXiv preprint arXiv:2003.08925*.

[50] Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

[51] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

[52] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi Parallel Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[53] Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-anuvadak: Tackling multiway translation of indian languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1781–1787.

[54] Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh M. Shah, and Pushpak Bhattacharyya. 2014. Shata-anuvadak: Tackling multiway translation of indian languages. In *LREC*.

[55] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

[56] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

[57] Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, et al. 2019. The niutrans machine translation systems for wmt19. In *Proceedings*

*of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266.

[58] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[59] Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817*.

[60] Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. pages 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

[61] Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, et al. 2019. Overview of the 6th workshop on asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35.

[62] Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. Overview of the 4th workshop on asian translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1–54.

[63] Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*.

[64] Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

[65] 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*, pages 875–880.

[66] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.

[67] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL (Demonstrations)*.

[68] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9. Association for Computational Linguistics.

[69] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics.

[70] 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[71] Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2020. Odiencorp: Odia–english and odia-only corpus for machine translation. In *Smart Intelligent Computing and Applications*, pages 495–504. Springer.

[72] Jerin Philip, Vinay P. Namboodiri, and C. V. Jawahar. 2018. CVIT-MT Systems for WAT-2018. In *5th Workshop on Asian Translation (WAT2018)*.

[73] Jerin Philip, Vinay P. Namboodiri, and C.V. Jawahar. 2018. CVIT-MT systems for WAT-2018. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

[74] Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2019. A baseline neural machine translation system for indian languages. *arXiv preprint arXiv:1907.12437*.

[75] Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, pages 43–70.

[76] Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

[77] 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*, pages 401–409. Association for Computational Linguistics.

[78] K Pramod Sankar, Saurabh Pandey, and CV Jawahar. 2006. Text driven temporal segmentation of cricket videos. *ICVGIP*.

[79] Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. *EACL 2017*, page 157.

[80] 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.

[81] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

[82] Sandeep Saini and Vineet Sahula. 2018. Neural machine translation for english to hindi. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–6. IEEE.

[83] Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 228–234.

[84] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

[85] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

[86] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, volume 1, pages 86–96.

[87] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

[88] Rico Sennrich and Martin Volk. 2010. Mt-based sentence alignment for ocr-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*.

[89] Rico Sennrich and Martin Volk. 2011. Iterative, mt-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182.

[90] Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

[91] Karanveer Singh and Pushpak Bhattacharyya. 2019. Nmt in low resource scenario : A case study in indian languages.

[92] Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2017. Comparing recurrent and convolutional architectures for english-hindi neural machine translation. *WAT*.

[93] Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

[94] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

[95] Themos Stafylakis and Georgios Tzimiropoulos. 2018. Zero-shot keyword spotting for visual speech recognition in-the-wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–529.

[96] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[97] TDIL-Sampark. Sampark Translation System.

[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

[99] Boli Wang, Zhixing Tan, Jinming Hu, Yidong Chen, et al. 2017. Xmu neural machine translation systems for wat 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 95–98.

[100] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481.

[101] Ziming Zhang and Venkatesh Saligrama. 2015. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174.

[102] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.