Chapter 1

Introduction

Machine Translation (MT), the concept of computers automatically translating between languages, has been around for a long time. It first appeared around 60 years ago in the works of Warren Weaver [1]. This first proposal translating languages using computers relied on information theory, which was also used for code-breaking in World War *II*. However, the earliest translation systems was rule-based. In the early days, machine translation was mainly focused on translating between English to other languages and vice-versa. However, globalisation brought out the need to consider as many languages as possible to include in the machine translation systems. MT is an important tool that can collect knowledge from other languages and distribute knowledge from one language to another.

In recent years, MT systems have been used for commercial and personal purposes. Companies use MT systems for commercial purposes to translate from English to various other foreign languages and vice versa. Microsoft, for example, used MT technology to translate technical manuals from English into various other languages. Many MT systems, such as Google Translate and Microsoft Bing Translator, are available online for public use.

Many works in MT focus on language-specific and language-independent aspects.

Language-specific aspects include modification in the MT in the form of language features, such as morphological, POS-tagging, orthographic and phonological information. In contrast, language-independent models include modification in the architecture of MT systems.

MT has to deal with various forms of linguistics irregularities in many languages simultaneously [2, 3]. It usually requires a massive parallel corpus to deal with irregularities in languages. *Corpus* is a language resource with a collection of large and unstructured text stored on the computer. MT achieves parity with human translation on resource-rich language pairs such as English-German and English-French [3]. Resource-rich or High Resource Languages (HRLs) are the languages for which a massive amount of parallel corpora are available to train the MT model. However, there are many languages (e.g., Nepali, Marathi) for which sufficient training resources are not available, called Low Resource Languages *(LRLs)*. MT faces data scarcity issues in achieving an acceptable performance for LRLs. In this dissertation, we specifically discuss developing a good MT system for low-resource Indian languages without relying on human power to generate more parallel corpora. Parallel corpora are text in language pairs, one of which can be called a source language and the other a target language. Each sentence in one language contains a translation in the other language. The development of an MT system for LRLs (low-resource machine translation) faces numerous challenges and issues, which are discussed in the following section.

1.1 Key issues in low-resource machine translation

MT researchers face various challenges while developing MT systems for low-resource Indian languages. Some of these are discussed as follows:

1.1.1 Lack of annotated and labelled datasets

All Indian languages, except Hindi (Hindi↔English for MT), fall into the low-resource category. The process of creating a parallel corpus for LRLs is time-consuming and labour-intensive. It is one of the most severe problems that MT researchers face. Because India has 122 major languages and 1599 other languages (according to the 2001 Census of India), manually creating data for each language is not always feasible.

1.1.2 Supporting multiple dialects of a language

Informally, dialects are regional varieties of a language that differ from other regional varieties in vocabulary, grammar and pronunciation and form a single language with them. Like for other languages, many dialects of languages exist for a single Indian language. For example, Bhojpuri, Magahi and Maithili each have many dialects. These varieties of languages pose a challenge in creating corpora for them.

1.1.3 Morphological richness

Morphologically Rich Languages (MRLs) are defined in the Computational Linguistics (CL)/Natural language Processing (NLP) literature as languages that express substantial grammatical information at the word level, such as information about the relationships between words or syntactic units, or about characteristics of the entities mentioned in the sentences. Morphological richness in languages increases the complexity of language models and hurts MT performance.

1.1.4 Word order

Although word order in Indian languages is relatively free, there are some units that occur in a fixed order. The most common examples are the main verb followed by auxiliary verb sequences and nouns followed by postpositions. The free word order nature of Indian languages poses different kind of challenges when compared with fixed

word order languages such as English.

1.1.5 Globalization pressure

Globally, using a few high-resource languages, such as English and German, can potentially lead to the extinction of other languages due to a lack of or death of native language speakers. Such languages are referred to as Endangered languages. Such languages can be saved with the help of MT. For example, Hebrew and Gaelic are two extinct languages that have been resurrected. MT help in survival or resurrection of languages, but MT is more difficult for such languages, as it is more challenging to create data and MT systems for low resource languages.

1.2 Motivation

This section of the dissertation discusses the motivation for developing MT tools and applications for low-resource languages.

1.2.1 Language preservation

The effort to prevent languages from becoming extinct is known as language preservation. When a language is no longer taught to younger generations and its fluent speakers (usually the elderly) die, it is on the verge of extinction. Written and spoken words are an art form, a means of passing down values and traditions to future generations. When a language is lost, a piece of culture is also lost. Similarly, when a language is preserved, the traditions and customs continue living in the hearts and minds of those who understand it. Language is more than just sentence structure and grammar; it is history, discourse, customs, and heritage. The preservation of a language implies the preservation of a culture. Therefore, MT acts as one of the essential tools for preserving the language and the culture.

1.2.2 Educational applications

Developing tools for translation between low-resource and high-resource languages can play an important role in improving the Indian education system. Learning in their native language, rather than in an unknown or a second language such as English, is more beneficial and easier for students. Instead, taking the time to translate the content from English to the native language by the human mind saves a lot of students' time and improves the students' learning capability. Sometimes even languages that were doomed to become extinct come back to thrive. Everyone knows the revival of Hebrew or Gaelic, and with the help of NLP techniques, such revival can be sped up drastically.

1.2.3 Knowledge expansion

Much of the world's knowledge is not contained in text; corpora records what people say but not what they mean, how they understood things, or what they did in response to the language. New developments in NLP may provide insights into the relationship between words and meaning — not through pure statistics but by comparing more diverse languages. MT systems can make better use of hidden knowledge stored in various languages.

1.2.4 Monitoring demographic and political processes

People who speak minor languages are usually hidden from our sight. However, when we consider that India has a population of over 1.39 billion people, we realise how important it is to get closer to them.

1.2.5 Emergency response

We are mankind, and we are equal in the eyes of God. Extending our prevention network with messages that more people can understand will save lives.

1.3 Research objectives

The research reported in this thesis focuses on extending the use of existing deep and reinforcement learning approaches to exploit the multilingual representation of languages in low-resource MT. The knowledge from various resources is combined with the least amount of annotated data from LRLs to produce cutting-edge results, indicating a promising research direction in MT. This thesis investigated the following four research problems, all of which focused on the processing of low-resource languages in order to improve MT systems:

1. Zero-shot problem: The problems faced by MTs become more challenging when the availability of training data is almost none or zero. We call such kinds of issues a Zero-Resource Problem (ZRP).
2. Morphological richness: Rich morphology poses different kinds of challenges in developing MT systems for low resource languages. We try to find ways to address these challenges.
3. Domain shift problem: The mismatch between training and test domain data causes domain shift, which degrades MT quality. We try to address this problem.
4. Missing-context and rare-word problem: Due to insufficient data, NMT fails to learn the proper context and it also faces the rare-word problem. This problem occurs when NMT fails to translate out-of-vocabulary or rare words not present in a training corpus. We try to address this by using sub-word representations.

In this thesis, we attempt to address the above-mentioned five problems by exploiting multilingual information between different low-resource related languages using various learning techniques such as projection of representations into a common space, transfer learning, domain adaptation, joint embeddings, pseudo-corpus generation and adversarial learning [4, 5]. Here, we can say "multilingual information" means semantic information represented in multiple languages. We have used transfer learning and domain adaptation techniques as a machine learning method where a model developed for

a one-language pair is reused as the starting point for a model of a related language pair. Moreover, we can say, "Transfer learning and domain adaptation refer to the situation where what has been learned from one language is exploited to improve generalisation to another related language". With the help of monolingual corpus, pseudo-corpus generation is used to create the synthetic parallel corpus in multiple languages. In adversarial learning, models generate translated samples using a generator and attempt to distinguish between real and translated samples in order to determine whether the generated sentences are close to fake or real.

1.4 Contributions

This section summarises the significant contributions and organisation of the thesis (Fig. 1.1). The key contributions of the thesis include the design, development, implementation, and comparative analysis of the proposed methods for addressing the problems faced by Indian low-resource MT. Chapters 2 and 7 describe the literature study and the thesis's conclusion, chapters 3 to 6 primarily include the major contributions of the thesis. The contributions and organisation of the thesis are as follows:

In chapter 2, we discuss some theoretical background and a literature review on MT's issues. The literature review is categorised into five sub-problems: zero-shot, morphological complexity, rare word, domain shift, and generalisation. We have proposed a solution to address to some extent each of these sub-problem in the following five chapters. This chapter also discussed the evaluation metrics used to perform comparisons between the proposed methods and the existing baseline approaches.

In chapter 3, we propose an approach based on leveraging the features of similar languages by simply, programmatically[1], converting them into an intermediate Latin-based multilingual notation. The notation that we use here is the commonly used WX-notation [6], which is often used in NLP tools and systems for Indian languages

[1]Using encoding converters, such as `https://pypi.org/project/wxconv/`

developed in India. This notation (like many other similar notations) can project all the Indic or Brahmi origin scripts [7], which have — in many cases — different Unicode blocks, into a common character space. Our intuition, is that this should help in capturing phonological, orthographic, and, to some extent, morphosyntactic similarities that will help a neural network-based model in better multilingual learning and translation across these languages [4, 5, 8, 9]. We do this by using this WX-converted text to learn byte pair encoding-based embeddings.

In chapter 4, we propose a Transfer Learning-based Semi-supervised Pseudo-corpus Generation (TLSPG) approach for zero-shot translation systems. TLSPG generates the pseudo corpus by exploiting the relatedness between low-resource and zero-resource language pairs via a transfer learning approach. It outperformed the state-of-the-art models by up to +15.56 on Bhojpuri→Hindi, +8.13 on Magahi→Hindi, +3.98 on Hindi→Bhojpuri and +2 on Hindi→Magahi language pairs, respectively. These four languages used in the experiments are closely related, which is favourable for training the MT in zero-shot scenarios. It is further empirically ascertained in our experiments that such relatedness helps improve the performance of zero-shot MT systems.

In chapter 5, we propose a REINFORCE-based Sentence Selection and Weighting (RSSW) method for filtering unrelated sentences from out-of-domain corpora. Many low-resource language pairs such as Hindi–Nepali and Spanish–Portuguese, for which sufficient parallel data are not available to train the model, face in-domain data scarcity issues. Non-availability of in-domain training data leads to the use of unrelated corpus in training the model, which degrades the NMT performance. This domain mismatch between training and test data results in a domain shift problem. To tackle the challenge of the domain shift problem in NMT, we propose the RSSW method, which selects pseudo-in-domain sentences from out-of-domain data and learns their weights based on Reinforcement Learning. In contrast to training on a combination of in-domain and out-of-domain data, filtering out-of-domain data creates pseudo-data near to the

in-domain, and training the model on this combination results in better translations.

In chapter 6, we exploit the phonological and morphological features to improve the performance of neural machine translation. The proposed model is based on the GAN model, where Deep-reinforcement guided attention model is used as a generator and the convolutional neural network as a discriminator. We also create the novel joint embedding of subwords and phonetic representations of sentences as input to GAN that helps models to learn better representations and generate suitable context vectors compared to existing traditional approaches for LRLs. The proposed model gains a considerable improvement on all language pairs. Our approach incorporates both sub-phonetic and subword embedding in GAN-NMT, achieving the best performance among all baselines. Another reason for the improvement is optimized attention in GAN-NMT, which leads to learning better context vectors.

Finally, in chapter 7, we summarise this thesis's contributions by drawing a conclusion and discussing future scope with challenges of the proposed approaches.

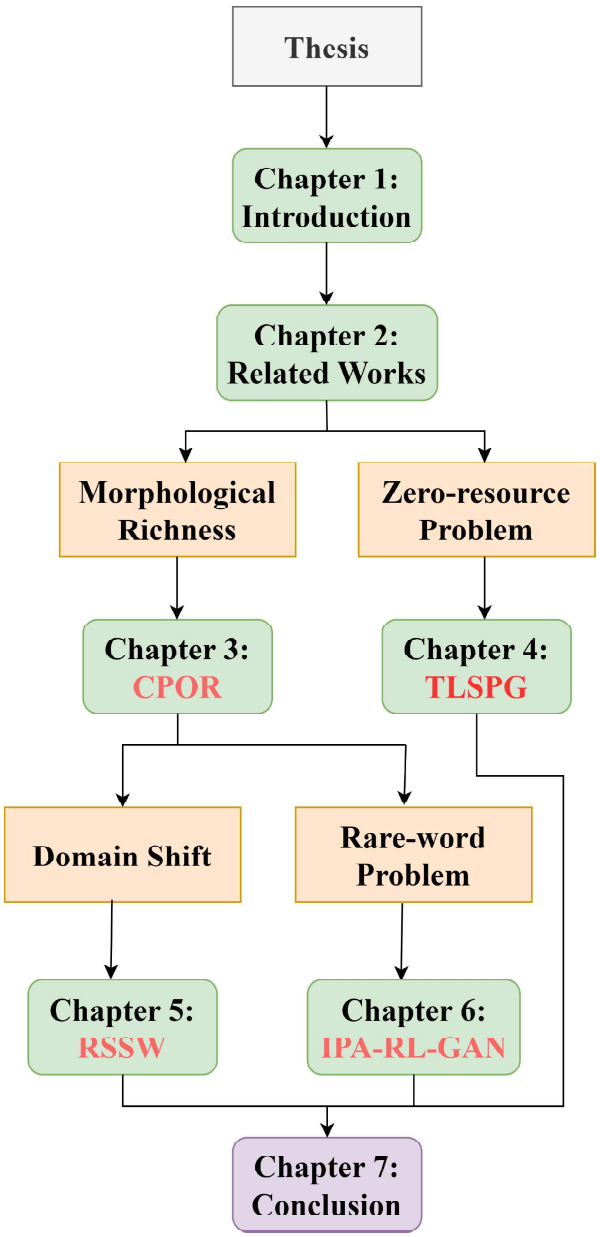

Figure 1.1: Thesis organization.

CPOR: Common phonetic-orthographic representation for NMT
TLSPG: Transfer learning-based semi-supervised pseudo-corpus generation approach for zero-shot NMT
RSSW: REINFORCE-based sentence selection and weighting method for NMT
IPA-RL-GAN: Incorporating IPA and reinforcement learning in GAN-NMT