Low Resource Similar Language Machine Translation in Common Phonetic Space with Multilingual, Adversarial and Reinforced Deep Learning

Thesis submitted in partial fulfillment
for the Award of Degree

Doctor of Philosophy

by

Amit Kumar

अमित कुमार

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. 17071018

December 2022

Chapter 7

Conclusions and Future Work

The goal of this dissertation is to create machine translation for morphologically-rich Indian languages that fall into low-resource categories and share orthographic, phonetic features to some extent, which can be used for exploiting similarity at the morphosyntactic level, in addition to orthographic and phonological levels. We carefully examined state-of-the-art MT system outputs and identified challenges that currently impede low-resource MT systems. Existing machine translation models struggle to train on Indian language data due to a lack of digitised resources, morphological richness, word order, and syntactic structures.

Table 7.1: Overview of the proposed approaches

Chapter	Proposed Method	Problem	Language Pair
3	CPOR	Morphological Richness	GU↔HI, NE↔HI, MR↔HI, PA↔HI, MAI↔HI, UR↔HI
4	TLSPG	Zero-resource Problem	BHO↔HI, MAG↔HI
5	RSSW	Domain Shift	NE↔HI, MR↔HI
6	IPA-RL-GAN	Rare-word Problem	GU↔HI, NE↔HI, PA↔HI, MAI↔HI, UR↔HI

We categorised the dissertation into four sub-problems: zero-shot, morphological complexity, rare-word and domain shift problems, and proposed solutions for each of them. The proposed solution architectures range from traditional SMT and NMT

Table 7.2: Summarization of Chapter-3

Model	SL	RR	Statistical	Neural	WX	Conclusion
Guzmán et.al [58]	✓	✗	✗	✓	✗	The proposed approach leads to improvement over baseline approaches by *0.6* BLEU points to *11.75* BLEU points.
Proposed approach	✓	✓	✗	✓	✓	

Note- SL: Similar Language, RR: Reducing Redundancy

Table 7.3: Summarization of Chapter-4

Model	Types of MT		Techniques		Training model	Conclusion
	Bilingual	Multilingual	Finetuning	Pivot	Neural	
mBART	✗	✓	✗	✗	Transformer	TLSPG outperformed the baseline models with +15.56 on BHO→HI, +8.13 on MAG→HI, +3.98 on HI→BHO and +2 on HI→MAG, respectively.
Kumar et al. [27]	✓	✗	✗	✗	Transformer	
TLSPG	✓	✗	✗	✗	Transformer	

systems to Reinforcement learning-based approaches. Since our main focus was on deep learning and reinforcement learning, the solutions are collectively referred to as deep-RL-based approaches. Table 7.1 contains the overview of the proposed approaches for different problems faced by low-resource Indian MT systems.

Our first solution (Table 7.2) is based on leveraging language relatedness by representing text in a shared phonetic-orthographic space. The second (Table 7.3) is based on a semi-supervised transfer learning approach for zero-resource languages, which takes advantage of the relatedness between zero-resource and low-resource languages. In the third solution (Table 7.4), we discussed domain shift and proposed a method for filtering related data from out-of-domain corpora using Reinforcement learning. Our fourth

Table 7.4: Summarization of Chapter-5

Model	Neural	REINFORCE	MLE	MRT	Conclusion
In-domain	✓	✗	✓	✗	Proposed method outperformed the existing approaches by ∼2 BLEU points for both language pair translation tasks.
MDA	✓	✗	✓	✗	
FTA	✓	✗	✓	✗	
RSSW + MDA	✓	✗	✓	✓	
RSSW + FTA	✓	✗	✓	✗	
Wang et.al [54]	✓	✗	✓	✗	
RSSW + MRT	✓	✓	✓	✓	

Note- MRT: Minimum Risk Training, MLE: Maximum Likelihood Estimation

Table 7.5: Summarization of Chapter-6

Model	SWE	SPE	IPA	BS	CTP	Conclusion
Transformer-NMT + PE	✓	✗	✓	✗	✗	Proposed approach suppress the existing techniques with +2 BLEU point s under both bilingual and multilingual settings
Transformer-NMT + GAN	✓	✗	✓	✗	✗	
Transformer-NMT + DRGA	✓	✗	✓	✗	✗	
Transformer-NMT + JE (Proposed)	✓	✓	✓	✓	✓	
Transformer-NMT + GAN + DRGA +JE (Proposed)	✓	✓	✓	✓	✓	

Note- SWE: Orthographic Sub-Word Embedding, SPE: Phonetic Sub-Word Embedding, BS: Beam Search, IPA: International Phonetic Alphabet, CTP: Sentence-wise BLEU comparison between predicted textual and phonetic sequences.

solution (Table 7.5) contributes by combining phonetic sub-word embedding with the orthographic sub-word embedding of morphologically rich and phonetic-similar languages.

The proposed solutions exploits multilingual information between different related Indian languages by progressing from traditional MT approaches to deep learning and from deep learning to reinforcement learning. Although SMT based approaches give better results for low resource languages if some parallel corpus is available, our main aim was to work on Zero Resource Setting and improve results by using Deep and Reinforcement Learning. Our proposed solutions make better use of the morphological, phonological, and relatedness features of low-resource languages for MT. We have used BLEU, chrF2 and TER as evaluation metrics to better judge the performance of our models.

In the future, we would like to investigate the capabilities of our MT models for other morphological and low-resource languages in greater depth. Although this thesis covers a few approaches for improving the translation quality of Indian low-resource MT, there are a lot of possibilities for future challenges encountered by the current MT system, some of which are discussed as follows:

1. In day to day real-life, people communicate in a code-mixed way in many places, including India. India is a multilingual country, and almost everyone uses at least two languages for communication. A lot of available data on the Web contains code-mixed data. Our approach mainly builds on representing data in a common phonetic-orthographic space using a WX or IPA. It is more challenging to use such

data, paricularly for low-resource languages. This is one of the major challenges we will try to address in the future.

2. The work described in the thesis can play an important role in MT for other non-Indian low-resource languages by generalising it to such language pairs which are closely related to each other (e.g., Spanish↔Portuguese). If common transliteration notation like IPA exists for other languages, the proposed approach can be generalised to non-Indian low-resource languages.
3. Big-tech companies like Meta and Microsoft are putting more of an emphasis on low-resource languages. They increasingly deploy powerful (and costly) methods and data for low-resource languages. Massive multi-NMT models are a promising technique, especially for systems produced by tech companies. In particular, multi-NMT for zero-shot translation is an important line of research, as it eliminates the need for parallel datasets between every possible language. However, all the early multi-NMT systems covered at most around 100 languages. Google's recent multi-NMT model with about 1,000 languages sets a new milestone in multi-NMT models with high language capacity. It provides reasonable performance in zero-shot cases by using monolingual data of LRLs. However, the question is how many researchers can afford to build such large models. This is partly due to the unavailability of parallel data, or even sometimes monolingual data. The work done in this thesis can play an important role in long-term improving the translation quality of low-resource and zero-resource languages. Our work is straightforward and advantageous for creating pseudo-data and raising translation quality at a lower cost.
4. We also want to be more effective in low-resource scenarios, where we can use cross-lingual information from other RL techniques to provide more data to greedy neural translation models.
5. Since our RSSW method is proposed for domain shift issues, we can broaden our

RSSW approach to the healthcare domain for local people who are more comfortable communicating in their local languages. Finding a professional interpreter with communication knowledge of less-resource languages is difficult. In a clinical encounter, a physician faced with a language barrier and no professional interpreter might choose to use a machine translator to assist in communicating with a patient. Physicians might also encourage patients to ask questions or respond to queries by directing them to input text into the machine translator in their language.

6. Similary, our RSSW approach can also be extended to the military and defence to communicate with local people during pandemics and save their lives. Communicating and creating awareness in the local language is more effective than communicating in the standard language.

The thesis describes some methods for creating and improving machine translation for Indian low-resource languages under low or zero resource scenarios. The proposed methods try to fill the research gaps of existing state-of-the-art methods. These works can not only help the society but the government organizations like healthcare and the defence.