Abstract

Neural Machine Translation (NMT) heavily depends on the context vectors generated via the attention network for the target word prediction. Existing works primarily focus on generating context vectors from words or subwords of sentences, limiting NMT models' ability to learn sufficient information about the source sentence representations. The situation is worse when languages belong to extremely low-resource categories. Although several approaches for low-resource Machine Translation (MT) have been proposed in the last decade, it accelerated after the introduction of deep learning, particularly after the Transformer model. Some of the significant challenges faced by low-resource MT are zero-resource problems, morphological richness, rare-word problems and domain shift issues. This dissertation attempts to address these problems by exploiting multilingual information between different low-resource languages using various learning techniques such as transfer learning, domain adaptation, joint embedding, Pseudo-corpus generation, adversarial and reinforcement learning.

Since most Indian languages use Indic or Brahmi origin scripts, they are highly phonetic in nature and similar in terms of abstract letters and their arrangements. As an initial step, we use these characteristics of Indian languages to tackle the morphological richness issue in low-resource MT and propose an approach based on common multilingual Latin-based encodings (WX notation) that takes advantage of orthographic and phonological similarities between languages. Text representation into common multilingual Latin-based encoding reduces the morphological complexity of languages

and improves the MT quality because WX does not distinguish between independent vowels (vowel akshars) and dependent vowels (maatraas). In the second solution, we solve the zero-resource problem in Indian languages by developing a Transfer Learning-based Semi-supervised Pseudo-corpus Generation (TLSPG) approach for zero-shot MT systems. It generates the pseudo corpus by exploiting the relatedness between low-resource and zero-resource language pairs via the transfer learning approach. It is further empirically ascertained in our experiments that such relatedness helps improve the performance of zero-shot MT systems.

Furthermore, we also tackle the challenge of the domain shift problem in NMT by proposing the third solution known as the REINFORCE-based Sentence Selection and Weighting (RSSW) method, which selects pseudo-in-domain sentences from out-of-domain data and learns their weights based on reinforcement learning. The proposed method leverages the similarity between language pairs by encoding the source and target languages into a common encoding script for language model training. RSSW uses minimum risk training and maximum likelihood estimation as an objective function to train NMT on selected pseudo-in-domain sentences. Moreover, in order to improve the learning of source sentence representations and handle the rare word problem of low-resource MTs, we propose an improvement in Generative Adversarial Networks (GAN)-NMT by incorporating deep-reinforcement learning-based optimised attention in generator and Convolutional Neural Network (CNN) in discriminator. This deep-reinforcement learning-based generator generates translations based on source sentences, and the CNN-based discriminator distinguishes whether a generated translation sentence is real or fake, i.e., whether machine-generated translation sentences are close to human-translated sentences. We also create the novel joint embedding of orthographic sub-word and phonetic sub-word representation of sentences as input to GAN that helps models to learn better representations and generate suitable context vectors compared to existing traditional approaches for low-resource NMT. Since we perform

experiments mainly on Indian languages, which are morphologically rich, feeding multiple morphologically-rich languages into a single model reduces the NMT's performance during training and affects the generalisation of models for multiple languages. We also apply all the proposed solutions on zero-resource language pairs and observe satisfactory improvement in translation qualities.

One the whole, the proposed solutions in this thesis exploit the multilingual information from different related Indian languages by progressing from traditional MT approaches to deep learning and from deep learning to reinforcement learning. BiLingual Evaluation Understudy (BLEU), character n-gram F-score-2 (chrF2) and Translation Edit Rate (TER) are used as evaluation metrics to judge the translation qualities produced by the different proposed models. Performance evaluation of the proposed approaches on Indian language pairs shows the effectiveness of the techniques over existing strong baseline methods.

Keywords: *Machine Translation, Low Resource Languages, Related Languages, Orthographic Information, Phonological Information*