Abstract

Machine Translation (MT) is one of the most essential applications of natural language processing (NLP) today. Current MT systems provide remarkable translation results for high resource language pairs, e.g., German and English, yet their effectiveness strongly relies on the availability of immense amounts of parallel data in the order of tens of millions of sentences. However, such data is not widely available for many low resource language pairs, that cause a bottleneck for MT. Therefore, improving MT on low-resource language pairs becomes one of the essential tasks in MT currently.

The aim of this study is to obtain an effective method for improving the translation quality of low-resource MT system on Myanmar-English language pair. It is worth nothing that Myanmar is a true resource-poor language and only small amount of Myanmar-English parallel sentence pairs are currently available to build baseline MT systems. Moreover, current Myanmar word segmentation tools may probably produce massive rare-words in both statistical MT (SMT) and neural MT (NMT) systems. Poor word segmentation can give an out-of-vocabulary (OOV) problem that will negatively impact to the overall translation performance. To this end, the main research question is as follows: how to employ deep learning architectures to improve MT for low-resource language pairs, dealing with the challenges of massive rare-words and data sparsity problem while having access to a limited amount of parallel data.

To answer the research question, we propose a framework with three subtasks as follows:

- Optimizing Myanmar Word Segmentation: puts a concentration on the performance of Myanmar word segmentation. Proper word segmentation can reduce massive rare-words and effective to the overall translation performance. Myanmar is a low-resource language; thus, only corpus-based, dictionary-based, rule-based, and statistical word segmentation tools are now available for free. After we analyze the results of these available tools, we address that these tools produce massive rare words in the MT tasks. Therefore, we propose a new unsupervised Myanmar word segmentation model [Zin et al., 2021] by learning only on the available MT corpora to produce the best segmentation results that is suitable for the later MT process. We observe that our proposed segmentation model can strongly support the Myanmar-English MT to achieve better translation performance against previous approaches.

- Automatic Post-Editing (APE): aims to correct systematic errors in a machine-translated text. We propose the three sub-modules for APE system including (i) word alignment information extraction, (ii) sentence enrichment, and (iii) sentence denoising. The proposed APE system can be effectively use as a post-processor to the existing low-resource NMT system for final editing the raw translated texts to meet the required quality standards. The first module conducts word alignments to detect the errors in a

machine-translated text. The second module enrich the input sentence by removing the detected errors and adding the missing information. The final module denoise the sentence to transform it into an accurate and fluent sentence.

- Creating Reliable English-Myanmar Parallel Corpus: deals with the parallel corpus expansion to ensure that the training corpus contains sufficient content. Two effective frameworks: *Construct-Extract* and *Expand-Extract* are proposed for creating the high-quality parallel corpora that are used to expand the training data. The *Construct-Extract* framework first constructs a pseudo parallel corpus from the target monolingual texts using back-translation (i.e., target-to-source translation) approach and then filters the noisy sentence pairs from the constructed corpus through a Siamese BERT-Networks based filtering system. The *Expand-Extract* framework first expands the existing comparable corpora using self-training (i.e., source-to-target translation) and back-translation, and then extracts only the high-quality sentence pairs from the expanded comparable corpora based on LaBSE sentence embeddings and an effective scoring method. We evaluate the effectiveness of our created parallel corpora by conducting the MT experiments. In the English-Myanmar low-resource scenarios, both *Construct-Extract* and *Expand-Extract* frameworks are useful for obtaining a reliable parallel corpus.

According to the results, each sub-task in the above obtains significant improvement compared to the previous works on Myanmar-English MT task. Further research could be undertaken to create the large-scale comparable corpora where huge-amount of parallel data can be augmented for additional training data and apply reinforcement learning for target-to-target text style transfer in the APE task.

Keywords: Myanmar Word Segmentation, Parallel Corpus Creation, Siamese-BERT Network, LaBSE Sentence Embeddings, Automatic Post-editing