Improving Low-Resource Machine Translation: A Case Study for Myanmar-English Pair

MAY MYO ZIN

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Improving Low-Resource Machine Translation: A Case Study for Myanmar-English Pair

MAY MYO ZIN

Supervisor ： NGUYEN Le Minh

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science
September, 2022

Doctoral Dissertation

Improving Low-Resource Machine Translation: A case study for Myanmar-English Pair

May Myo Zin

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Science (Information Science)

Written under the direction of
Processor Minh Le Nguyen

September, 2022

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

Acknowledgments

First of all, I would like to express my sincere gratitude to my principal advisor, Professor Nguyen Le Minh of Japan Advanced Institute of Science and Technology (JAIST), for his constant encouragement, continuous support and kind guidance during my Ph.D. study. He has gently inspired me in researching as well as patiently taught me to be strong and self-confident in my study. Without his consistent support, invaluable patience and kindness I could not finish the work in this dissertation.

I would like to thank my second supervisor Professor Satoshi Tojo, my sub-theme advisor Associate Professor Kiyoaki Shirai, and Associate Professor Okada Shogo of JAIST, and Professor Ittoo Ashwin of Liège University for useful discussions and comments and generous support on this dissertation.

A special thanks to Assistant Professor Teeradaj Racharak of JAIST. I have received a lot of help from him. He gave me invaluable advice, comments, and most importantly provided emotional support to me all the time.

I am deeply indebted to the Ministry of Education, Culture, Sports, Science and Technology (MEXT) for granting me a scholarship during the period of my research. Thanks also to the Asian Office of Aerospace Research and Development (AOARD) for providing me with their Air Forced Office of Scientific Research (Grant no. FA2386-19-1-4041) which supported me to attend and present my work at international conference.

I would like to thank JAIST staff for creating a wonderful environment for both research and life. I would love to devote my sincere thanks and appreciation to my tutor Dr. Nguyen Tien Huy and all members of Nguyens laboratory. Being a member of Nguyens lab and JAIST is a wonderful time of my research life. I must thank Myanmar friends in JAIST for sharing my happiness and difficulties and cheered me up all the time.

Last but not the least, I would like to thank all members of my family for supporting me with great patience and love. I would never complete this work without their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Machine Translation

In today's interconnected world, translation became an essential tool, allowing us to communicate and share information, no matter what the language. However, translating large amounts of content could bring complications around cost, quality, and time. Therefore, machine and technology have come to our help in order to remedy some of these potential issues.

Machine Translation (MT) systems are applications that use machine-learning technologies to translate vast amounts of text from and to any of their supported languages. MT system automatically translates content form one natural language (the source) to another (the target), preserving the meaning of the source text, and producing fluent text in the target language. Although the concepts behind MT technology is relatively simple, the science and technologies behind it are extremely complex and bring together several leading-edge technologies, in particular, deep learning (artificial intelligence), big data and linguistics.

On looking into the history of MT, the most frequently used MT technologies are Rule-based MT (RBMT), Statistical Machine Translation (SMT) and Neural MT (NMT). RBMT usually generates consistent and stable translations based countless built-in linguistic rules and gigantic bilingual dictionaries for each language pair. Although RBMT can achieve good results, the training and development costs are very high for building a good quality system. As language is constantly changing, rules must be managed and updated where necessary in RBMT systems. Moreover, RBMT system sometimes results in a lack of fluency. Table 1.1 shows an example of an RBMT translation from German to English. The problem with this sample is that the grammar is not correct, and it doesn't sound fluent in the translation output.

Table 1.1: An example translation of RBMT translation from German to English

German → English	
Starten Sie die Wiedergabe am angeschlossenen Gerät und stellen Sie eine moderate Lautstärke ein.	Playback starts from the connected device and set a moderate volume.

In the nineties, the substantial increase in computer speed and storage capacity gave a way to the rise of SMT technology. SMT estimate the best possible translations for a word given the context of a few words based on statistical models derived from large amount of parallel corpus (a large and structured set of human-translated texts between two languages). SMT can practically handle any language pair where parallel corpus is present and its translation sounds much more fluent compared to RBMT as shown in Table 1.2.

Table 1.2: An example of RBMT translation from German to English

German → English	
Starten Sie die Wiedergabe am angeschlossenen Gerät und stellen Sie eine moderate Lautstärke ein.	Playback starts from the connected device and set a moderate volume.

As an emerging method of machine translation based on neural networks, Neural Machine Translation (NMT) caused a radical shift in translation technology, resulting in much higher quality translations. As shown in Table 1.3, NMT provides translations that are much more fluent and readable than RBMT and SMT.

Table 1.3: An example of RBMT translation from German to English

German → English	
Starten Sie die Wiedergabe am angeschlossenen Gerät und stellen Sie eine moderate Lautstärke ein.	Start playback on the connected device and set the volume to a moderate level.

Choosing the best MT technology may depend on the volume of training data and the type of content that we want to translate. RBMT is better for the translation of small content volumes and will provide consistent translation quality for a fixed set of terminology data and for short sentences. In comparison to RBMT, SMT and NMT are the most recent technologies that have made huge progress in terms of fluency and contextual accuracy and make most sense when we need to translate in high volumes such as technical manuals. However, the translation quality of both SMT and NMT can only be offered when a large amount of high-quality training data (up to millions of parallel sentences) is available.

1.2 Research Problems and Contributions

Current SMT and NMT systems achieve near human-level performance on some language pairs such as English-French and English-German, yet their effectiveness strongly relies on the availability of vast amounts of parallel corpora. The biggest issue with low-resource language pair such as Myanmar-English is the extreme difficulty of obtaining enough parallel corpus because paired sentences are usually expensive to create, as it requires an expert to translate sentence in one language into another language.

The main aim of this thesis is to improve existing MT performance for low-resource language pairs by effectively improving and integrating different component technologies of MT in an ideal or at least improved workflow. In order to achieve our objective, we propose three potential approaches that can improve the translation quality of existing low-resource MT systems, taking a challenging Myanmar-English pair for benchmarking. In this dissertation, I will show that with (1) optimization of Myanmar word segmentation, (2) training data augmentation, and (3) effective automatic post-editing, we can in fact build a competitive MT system with only limited data at hand.

For Myanmar, there is no distinct word boundary. Word segmentation is widely applied as a pre-processing step in MT pipelines. However, conventional Myanmar word segmenters probably produce massive rare-words/unseen-words, and most of the low-frequency words will be discarded during training. SMT systems cannot translate words that are unseen in the training corpus (known as out-of-vocabulary or OOV words) [Paul et al., 2008, Liu et al., 2018]. Dealing with morphologically rich languages like Myanmar and having a limited amount of bilingual resources makes this problem even more severe. The same problem also occurs in NMT system. Among the six challenges of NMT coined by [Koehn et al., 2017], OOV problem is considered the most serve one, especially in translation of low-resource languages. NMT system cannot make use of rare-words because it commonly limits the vocabulary to most high-frequency words [Wang et al., 2018]. Myanmar word segmentation is a necessary step in Myanmar-English SMT and NMT systems and its performance has an impact on the MT results. As a common solution in Myanmar-English translation, Myanmar texts are segmented using an off-the-shelf segmenter [Pa et al., 2008] which is trained on a manually segmented corpus. However, the domain of the segmented corpus may not exactly match with the MT task at hand. Consequently, the performance of the MT systems will be influenced. This leads us to our first research question (RQ):

RQ1: *How to optimize Myanmar word segmentation for improving MT performance?*

The ultimate goal of MT systems is to provide fully automatic agreed-level quality translations. However, existing MT systems often fail to produce the agreed level quality output. That is where MT with post-editing comes into play – a workflow where the raw MT output is corrected or post-edited by a human translator prior to final delivery (cf. Figure 1.1).

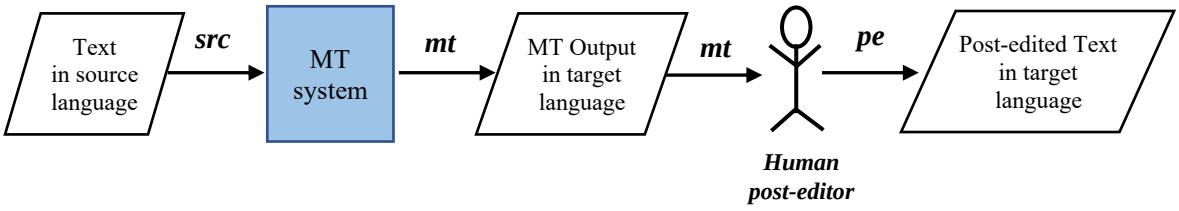

Figure 1.1: Post editing on MT output; *src*: source sentence, *mt*: corresponding MT output, and *pe*: human post-edited version of *mt.*

Automatic post-editing (APE) systems seek to perform the same task as human post-editors: correcting systematic errors in text produced by MT systems. APE systems have been used to correct different types of MT errors, from determiner selection [Knight et al., 1994] to grammatical agreement [Marecek et al., 2011]. However, APE task encounters a chronic problem concerning training data generation [Negri et al., 2018, Lee et al., 2021]. Generally, data for the APE system comprises the source sentence (*src*), machine translation of the sentence (*mt*), and corresponding human post-edit sentence (*pe*), collectively known as APE triplets. Building these data require extensive and expert-level human effort, as it contains an

elaborate process that involves identifying errors in sentences and providing suitable revisions. This occurs the absence APE training data for most language pairs and limits the effectiveness of APE task. This leads to our second research question:

RQ 2: *How can we build an effective APE system which can improve the translation quality of first-stage MT system, without having access to APE triplets?*

Large-scale parallel corpora are essential for training high-quality MT systems; however, such corpora are not freely available for many language pairs. In contrast, monolingual texts and comparable corpora are easier to obtain. Previously, training data has been augmented by either a pseudo-parallel corpus constructed from the target monolingual texts through the back-translation process or a corpus which contains nearly parallel sentences extracted from the comparable corpora. However, in the low-resource language pairs, in which only low-accurate machine translation systems can be used in back-translation process. Therefore, the constructed pseudo-parallel corpus may contain the noisy sentence pairs. The final translation quality may degrade when a low-quality pseudo-parallel corpus is naively used as the additional training dataset. To improve machine translation performance with low-resource language pairs, a method to effectively expand the training data via filtering the pseudo-parallel corpus is required. This motivated us to pose and explore our third research question:

RQ3: *How can a reliable parallel corpus be created from the target monolingual texts?*

Moreover, although the currently collected English-Myanmar comparable corpora contain large in-domain monolingual texts, only a small number of nearly parallel sentence pairs are contained in that corpora. Small amount of additional training data is not very effective to improve the performance of MT systems. Therefore, a method to better exploit the existing comparable corpora is required. This leads us to the fourth research question:

RQ4: *How can a reliable parallel corpus be created from the existing comparable corpora?*

All of the above research questions (RQ1, RQ2, RQ3 and RQ4) are answered through three subtasks, which are shown in Figure 1.2, as follows:

1. **Optimizing Myanmar Word Segmentation** considers the OOV problem. We attempt to analyze and tackle the problem of OOV words which frequently occur in the existing MT task of Myanmar-English low-resource language pair. Byte-pair-encoding (BPE) algorithm [Sennrich et al., 2015] is a simple data compression technique that replaces the most frequently occurring pair of bytes with a single, unused byte. In order to improve the performance of existing MT system, we specifically propose an unsupervised Myanmar word segmentation approach based on the NFKC (*Normalization* Form Compatibly Composition) normalization and BPE mechanisms. The proposed segmentation approach can learn itself to adapt the current MT domain and significantly reduce the OOV rate.

2. **Automatic Post-editing (APE)** aims to improve MT systems by automatically

correcting systematic errors in the MT results. The goals of APE task are (1) to cope with systematic errors of an MT system whose decoding process is not accessible, (2) to provide professional translators with improved MT output quality to reduce human post-editing effort, and (3) to adapt the output of a general-purpose system to the lexicon/style requested in a specific application domain. To develop the APE model, most studies require APE triplets including a source sentence (*src*), machine translation sentence (*mt*), and human post-edited sentence (*pe*). There is no APE triplets for Myanmar-English language pair because of low-resource. Instead of training on APE triplets, we develop an APE system on available monolingual and parallel data. The proposed APE system which is a pipeline consisting of three main modules: word alignment information retrieval, target sentence enrichment, and target sentence denoising; shows significant improvements on the quality of existing MT output.

3. **Building Reliable Parallel Corpus** puts the consideration on increasing the amount of training corpus. The quality of the MT systems depends on the availability of vast amount of parallel corpus. This is the major issue for low-resource language pairs where the size of available parallel corpus is too small; leads to poor translation quality. Findings in the literature show that there are two approaches that support the automatically creation of parallel corpus that can be supplemented to the training data.

    3.1 The first approach is back-translation that constructs the pseudo parallel corpus from the target monolingual texts. In low-resource language pairs, in which only low accuracy MT systems can be used in back-translation process, the constructed pseudo parallel corpus may contain low-quality sentence pairs. Data quality plays an essential role in training high-quality MT systems. The translation quality of MT system may degrade when a low-quality pseudo-parallel corpus is naively used as additional training data. To create a high-quality pseudo-parallel corpus from target monolingual texts, we propose a *Construct-Extract* framework that constructs pseudo-parallel corpus using back-translation approach and then extracts only high-quality sentence pairs from the constructed corpus using a Siamese BERT-Networks based approach.

    3.2 The second approach is parallel sentences extraction from comparable corpora. The extracted parallel sentences are used as the additional training corpus. English-Myanmar is a low-resource language pair and thus the collected comparable corpora contain only a small number of nearly parallel sentence pairs. Small amount of additional sentence pairs is not very effective to improve the performance of MT systems. To obtain more reliable parallel sentences from the existing comparable corpora, we propose an *Expand-Extract* framework that effectively expands the existing comparable corpora by augmentation source from existing target sentences and vice versa, and then extract only reliable sentence pairs from the expanded corpora.

The proposed frameworks can be the most useful approaches for obtaining a reliable parallel corpus in Myanmar-English low-resource scenario. MT systems trained using

our created parallel corpus as an additional data returned the best translation performance.

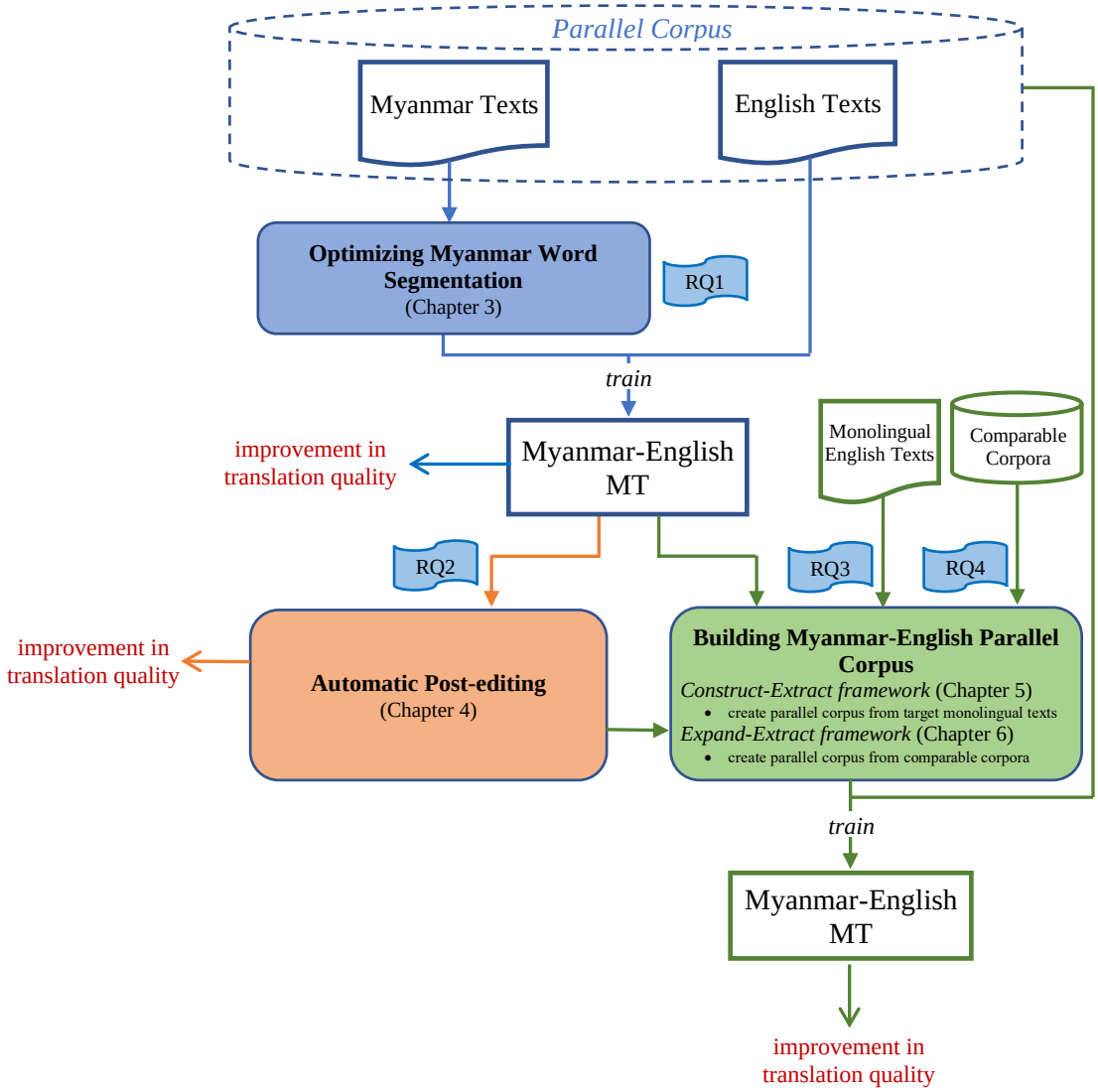

Figure 1.2: Schematic design of the research and the research questions presented in this thesis.

1.3 Dissertation Outline

The remainders of this thesis are organized as follows:

Chapter 2 introduces the background information about machine translation models and pre-trained embedding models as they are related to this thesis. Furthermore, we also describe the problem statements of machine translation systems in low-resource settings. Then, we present approaches towards improving the low-resource machine translation task.

Chapter 3 presents our study on Myanmar word segmentation for achieving highly accurate translations. We first describe the Myanmar word segmentation task and challenges. Next, we do a literature review to analyze the gaps in current methods. Then, we present our word

segmentation proposal by combining the idea of Unicode NFKC normalization and byte-pair encoding (BPE) mechanism. Finally, we describe our evaluation, experiments, and details of the discussion against traditional approach.

Chapter 4 explains our study on the integration of automatic post-editing (APE) into the machine translation system to ensure that the quality of the final translation meets desired quality standards (adequate and fluent). First, we analyze the raw outputs of the MT systems. There are two kinds of errors: incorrect words in target translation (semantic gap / low adequacy) and informal structure in target language translations (grammatical errors / low fluency). Through recent approaches, in automatic post-editing process, we present a simple and effective APE framework that can significantly increase the translation quality by providing the effectiveness of minimizing the semantic gap between the source and target sides and transferring the final output into the formal target style fluently.

Chapter 5 proposes the novel approach on the English-Myanmar bilingual corpus creation task from the collected English monolingual texts. In the literature review, we firstly survey some recent approaches for constructing additional pseudo-parallel data and extracting high-quality sentence pairs from the comparable corpora. Then, we investigate the effective framework of building high-quality bilingual corpus. The framework called Construct-Extract is composed of two subtasks: pseudo-parallel corpus construction and high-quality sentence pairs extraction.

Chapter 6 presents another effective approach for creating English-Myanmar parallel corpus from the collected comparable corpora. First, we do a literature review. Then we present our proposed framework called Expand-Extract which is composed of two subtasks: expanding the existing comparable corpora and extracting reliable parallel sentences from the expanded corpora. In experiments and discussion section, we describe experimental results and some qualitative analysis for making a conclusion.

Chapter 7 concludes our research and discusses future directions based on our works.

Chapter 2

Background

2.1 Machine Translation Framework

Currently there are two major machine translation approaches: Phrase Based Statistical Machine Translation (PBSMT) [Koehn et al., 2003] and Neural Machine Translation (NMT) [Vaswani et al., 2017]. Both approaches rely on parallel corpus, which contain parallel pairs of source language sentence $X = x_1, \dots, x_n$ of length n and its translation $Y = y_1, \dots, y_m$ of length m in target language. In the following subsections, we will provide the rough overview of SMT model (PBSMT) and NMT model (Transformer).

2.1.1 Statistical Machine Translation

Statistical machine translation (SMT) uses large amount of parallel corpus to find the most probable translation for a given input sentence. SMT systems learn to translate by analysing the statistical relationships between original texts and their existing human translations. The most important components in SMT are the translation model and the language model.

Translation Model:

SMT engine is trained with parallel corpus to create a translation model. A translation model is a table of aligned phrases and their translation. These phrases are called n-gram. The purpose of the translation model is to predict candidate translations for specific input texts. The translation model can be associated with adequacy because it preserves the meaning of the source.

Language Model:

The language model is built from the target language monolingual texts. The language model finds the best choice from the candidate translations based on the translation language. The language model can be associated with fluency in the translation because it gives the translated text its natural language flow.

The process of SMT system is as follows:

1. The input text (source sentence) is divided into phrases.
2. The phrases are matched with their parallel equivalents from the translation model.
3. The language model validates that the translation is probable in the target language.

There are four different SMT models: (1) word-based model which generates the translation word-by-word, (2) phrase-based model which translates sequences of words, (3) syntax-based

model which translates syntactic units, and (4) hierarchical phrase-based model which combines the strengths of phrase-based and syntax-based translation. Among these models, phrase-based SMT (PBSMT) is the most dominant paradigm. In this thesis, we use PBSMT as the baseline SMT system.

The basic PBSMT model is an instance of the noisy-channel approach [Brown et al., 1993], in which the translation of a Myanmar sentence X into an English sentence Y is modeled as follow:

$$argmax_Y\, P(Y|X) = argmax_Y\, P(X|Y)\, P(Y) \tag{2.1}$$

where $P(X|Y)$ is called the translation model which is trained on parallel corpus and $P(Y)$ is the language model trained on monolingual target language texts only. The translation model $P(X|Y)$ encodes Y into X by the following steps:

1. segment Y into phrases $\bar{y}_1 \ldots \bar{y}_m$, typically with a uniform distribution over segmentations;
2. reorder the $\bar{y}_i$ according to some distortion model;
3. translate each of English phrase $\bar{y}_i$ into a Myanmar phrase $\bar{x}_i$ according to a model $P(\bar{x} \mid \bar{y})$ estimated from the training corpus.

Most recently published methods on extracting a phrase translation table from a parallel corpus start with a word alignment. GIZA ++ is the most common tool to establish a word alignment. The decoder of PBSMT model employs a beam search algorithm similar to the one used by Jelinek (book "Statistical Methods for Speech Recognition", 1998) for speech recognition.

2.1.2 Neural Machine Translation

Most competitive neural machine translation (NMT) models have an encoder-decoder structure [Kalchbrenner et al., 2013; Sutskever et al., 2014; Bahdanau et al., 2015]. The encoder of NMT model maps an input sequence of symbol representations $(x_1, \ldots, x_n)$ to a sequence of continuous representations $z = (z_1, \ldots, z_n)$. Given z, the decoder of NMT model generates an output sequence $(y_1, \ldots, y_m)$ of symbols one element at a time. NMT model is auto-regressive [Graves 2013] at each step, consuming the previously generated symbols as additional input when generating the next. Recently, a novel deep neural network model, Transformer [Vaswani et al., 2017], with an innovative multi-head attention mechanism has been introduced. Transformer has become the state-of-the-art model for many artificial intelligence (AI) tasks, including MT [Zhang 2018; Tetko et al., 2020; Wolf 2020]. In comparison with other NMT models, including recurrent neural networks (RNNs), Transformer model not only provides better translation results but also can be trained in a shorter period of time. In this thesis, we used the transformer model as the baseline NMT system.

While we refer readers to [Vaswani er al., 2017] for a more complete formulation, we will briefly describe the overview of Transformer in this section. Like RNN models, Transformer

is designed to process sequential input data. However, unlike RNNs, the Transformer model does not necessarily process the input data in sequential order. Instead, the self-attention mechanism (shown in Figure 2.1) identifies the context which gives meaning to each position in the input sequence, allowing more parallelization than RNNs and reducing the training time.

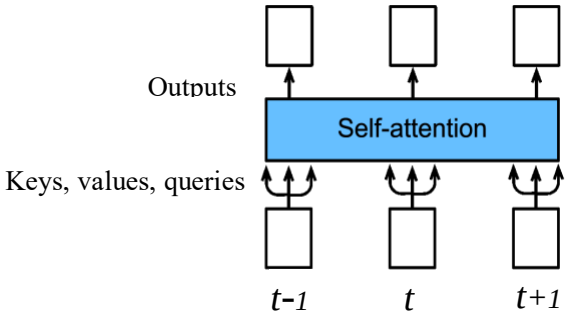

Figure 2.1: Self-attention mechanism

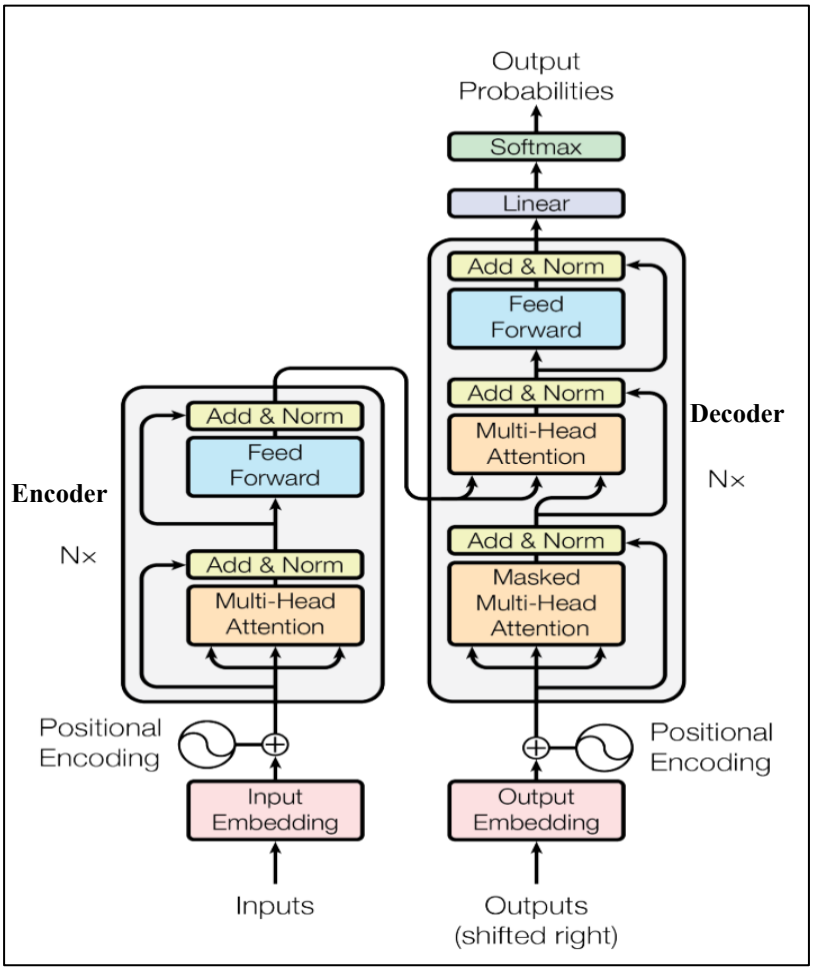

Figure 2.2: The Transformer – model architecture [Vaswani et al., 2017]

The model architecture of Transformer is shown in Figure 2.2. Transformer model is also based on the encoder-decoder architecture similar to the Sequence-to-Sequence (seq2seq) MT model. However, the Transformer model differs from the seq2seq model in the following three ways:

Transformer Block:

The recurrent layer of encoder and decoder in the seq2seq model is replaced by a Transformer Block. This block contains a multi-head self-attention layer and a position-wise fully connected feed-forward network layer for the encoder. Another multi-head attention layer is used to compute the encoder state for the decoder.

Add & Norm:

A residual connection [He eta al., 2016] is employed around each of the multi-head self-attention layer and position-wise fully connected feed-forward layer, followed by layer normalization [Ba et al., 2016].

Positional Encoding:

Since the Transformer model contains no recurrence and no convolution, in order for the model to make use of the order of items in a given sequence, some information about the relative or absolute position of the items in the sequence. To this end, positional encodings is added to the input embeddings at the bottoms of the encoder and decoder stacks.

2.2 Pre-trained Embedding Models

Contextualized word embedding models such as BERT [Devlin et al., 2019] and RoBERTa [Liuet al., 2019] represent words using continuous vectors calculated in context, and have achieved impressive performance on a various downstream of NLP tasks. Multilingual embedding models are powerful tools that map text from different languages to a shared vector space (or embedding space). This implies that in this embedding space, similar or related words will lie closer to each other, and unrelated words will be distant (cf. Figure 2.3). Multilingually trained word embedding models such as multilingual BERT (mBERT) [Devlin et al., 2019] can generate contextualized embeddings across different languages.

Existing approaches for generating sentence embeddings such as LASER [Schwenk et al., 2017] or MUSE [Conneau et al., 2018] rely on large amount of parallel data to a sentence form one language to another in order to encourage consistency between the sentence embeddings. These approaches yield good overall performance across a number of languages. However, they often underperform compared to dedicated bilingual models which use translation pairs as training data to obtain more closely aligned representations. Moreover, due to poor quality of training data (especially for low-resource languages) and limited model capacity, it can be difficult to extend these models to support a large number of languages while maintaining good performance.

Figure 2.3: Multilingual Embedding Space[1]

2.2.1 Language-Agnostic BERT Sentence Embedding

Language-agnostic BERT Sentence Embedding (LaBSE) [Feng et al., 2022] is currently the best method for parallel sentences mining. LaBSE is trained on 17 billion monolingual sentences and 6 billion bilingual sentence pairs using masked language modeling (MLM) and translation language modeling (TLM) [Conneau et al., 2019] pre-training. LaBSE is effective even on low-resource languages for which there is no data available during training process.

As illustrated in Figure 2.4, LaBSE model make use of dual-encoder models, which are an effective approach for learning bilingual sentence embeddings [Guo et al., 2018; Yang et al., 2019]. Embeddings of source and target sentences are extracted from each encoder. Then, the cross-lingual embeddings are trained using a translation ranking task with in-batch negative sampling:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{\phi(x_i,y_i)}}{e^{\phi(x_i,y_i)} + \sum_{n=1,n\neq i}^{N} e^{\phi(x_i,y_n)}} \tag{2.2}$$

Typically, the embedding space similarity of x and y is given by $\phi(x,y) = xy^T$. The objective is to maximize the compatibility between the source x_i and its true translation y_i and minimize it with others (negative sampling). As far as "bidirectional" is concerned; the final loss can sum the source-to-target loss $\mathcal{L}$ and target-to-source loss $\mathcal{L}'$ [Yang et al., 2019].

$$\bar{\mathcal{L}} = \mathcal{L} + \mathcal{L}' \tag{2.3}$$

To improve the separation between translations and nearby non-translations, LaBSE use additive margin softmax which extends the scoring function ϕ by introducing margin m around positive pairs [Yang et al., 2019]:

[1] https://ai.googleblog.com/2020/08/language-agnostic-bert-sentence.html

$$\phi'(x_i, y_j) = \begin{cases} \phi(x_i, y_j) - m & if\ i = j \\ \phi(x_i, y_j) & if\ i \neq j \end{cases} \tag{2.4}$$

Then the additive margin loss is as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \frac{e^{\phi(x_i, y_i) - m}}{e^{\phi(x_i, y_i) - m} + \sum_{n=1, n \neq i}^{N} e^{\phi(x_i, y_n)}} \tag{2.5}$$

LaBSE model establishes a new state-of-the-art on multiple parallel sentence pairs retrieval tasks. However, in this thesis, we investigates LaBSE model in parallel word pairs (word alignments) retrieval task.

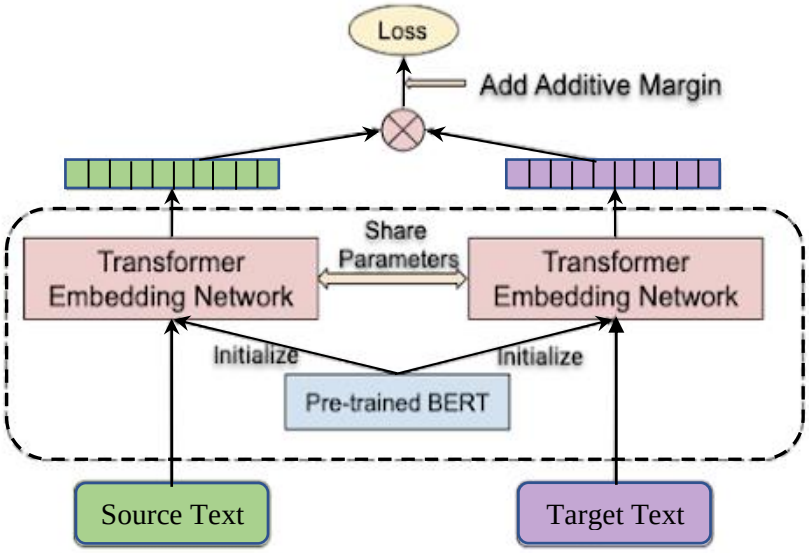

Figure 2.4: Dual encoder model with BERT based encoding modules.

2.3 Challenging Issue with Low-resource Languages

Most of the language pairs lack adequate amounts of parallel corpora, so existing SMT an NMT systems have difficulty producing accurate translations for these language pairs. This research focus on Myanmar-English low-resource machine translation where we have only 224,634 parallel sentences.

In the following sections, we will describe the challenges and potential solutions of existing Myanmar-English MT systems in low-resource scenarios.

2.3.1 Out-of-Vocabulary

An important feature of Myanmar texts is that they do not explicitly indicate word boundaries by spaces. Word segmentation is an important first step in MT, and its performance has a great impact on MT performance. NMT is primarily based upon word-level with a fixed vocabulary. Therefore, low resource morphologically rich languages such as Myanmar are mostly affected by the out-of-vocabulary (OOV) words and Rare word problems. On the other hand, Phrase-based SMT (PBSMT) system train its statistical model using parallel corpus. For unknown words, no translation entry is available in the statistical translation model (phrase-table). Therefore, words that do not appear in the training corpus cannot be translated by SMT system. Dealing with languages of rich morphology like Myanmar and having a limited amount of bilingual resources makes this problem even more severe.

In general, the causes of OOV words can be mainly categorized into the following. OOVwords result from segmentation error. Another source of OOV problem can be attributed to name entity such as person, location and organization. Finally, OOVs may originate from low-frequency abbreviations and combination forms of common words.

Adapting or optimizing word segmentation for MT systems can reduce the OOV rate [Chang et al., 2008; Paul et al., 2008; Wang et al., 2018]. Most of the current off-the-shelf segmenters trained on a manually segmented corpus. Therefore, the domain of the segmented corpus may not exactly match with the MT task at hand. Consequently, the disambiguation ability of the segmenter will drop and the performance of the MT systems will be influenced. To optimize Myanmar word segmentation that perfectly adapt the domain of the MT task at hand, we design unsupervised segmentation model by leveraging NFKC normalization and BPE mechanism.

2.3.2 Systematic Translation Errors

MT systems aim to provide fully automatic publishable quality translations. Although NMT systems show very promising results for some language pairs, their output still needs to be post-edited by human translators. [Daems et al., 2017] reported machine translation error types with the greatest impact on post-editing effort. The final classification used in their study can be seen in Figure 2.5 where the most common MT errors overall are grammatical errors (grammar and syntax), followed closely by adequacy issues. From the perspective of ‘acceptability,’ it was the grammar and syntax category, which turned out to be the most common error category for MT output, with word order issues, structural issues, and incorrect verb forms occurring more than 10 times each. From the perspective of ‘adequacy,’ other meaning shifts and word sense issues occurred frequently enough to be considered as separate categories, while the other subcategories (additions, deletions, misplaced words, function words, part of speech, and inconsistent terminology) were grouped together into ‘adequacy other.’

Automatic post-editing (APE) aims to improve the quality of machine translations, thereby reducing human post-editing effort by automatically fixing errors in the machine-translated text. However, the training of APE models has been heavily reliant on vast amount of artificial corpora combined with only limited human post-edited data. This is a main issue for low-

resource language pairs including Myanmar-English where APE data is still unavailable. In order to overcome this limitation, we investigate a simple and effective APE model by leveraging pre-trained embedding model and denoising mechanism using only available monolingual and parallel data, without having access to APE data.

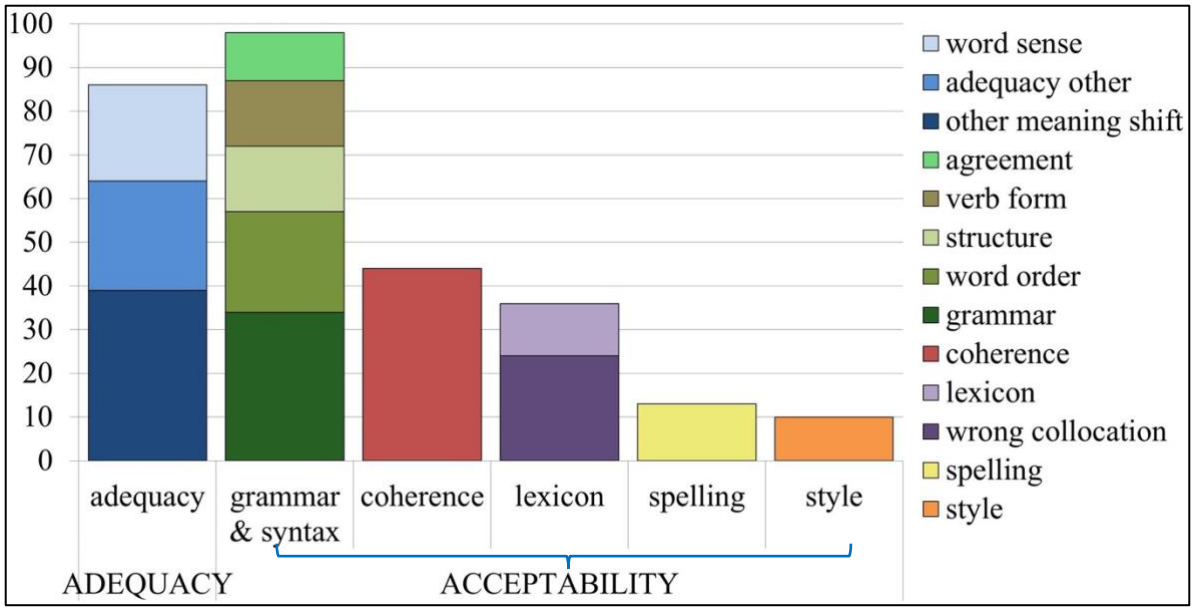

Figure 2.5: Error type frequency in MT output. [Daems et al., 2017]

2.3.3 Data Sparsity Problem

A large-scale parallel corpus is an essential resource for training highly accurate SMT and NMT systems. Manual creation of a high-quality and large-scale parallel corpus requires time, financial resources, and human experts for translating a large amount of text. Therefore, most of the language pairs have no or less parallel corpora. Due to the lack of sufficient parallel corpus, leveraging data other than parallel corpus is essential in low-resource MT systems.

Since collecting monolingual texts and comparable corpora are much easier and of lower cost than parallel corpus. Various approaches have been proposed to automatically create a pseudo-parallel corpus by leveraging monolingual data. For example, Zhang et al., [2016] proposed a method to obtain more pseudo-parallel data by translating source-side monolingual sentences with the baseline MT trained on available parallel data. Sennrich et al., [2016] achieved substantial improvements by automatically translating target monolingual texts into source language to obtain pseudo-parallel corpus and using the obtained corpus as an additional training data. However, they experimented only on the language pairs for which relatively large-scale parallel corpora are available. They did not fully exploit the training corpus or address the quality of the pseudo-parallel corpus. The quality of pseudo-parallel corpus is critical. Using low-quality parallel sentences will degrade NMT performance more than SMT [Koehn et al., 2017]. Accordingly, our first motivation related to data sparsity problem is to

automatically create pseudo-parallel corpus by leveraging English monolingual data and filter out low-quality synthetic sentences that might be included in such a pseudo-parallel corpus for obtaining a high-quality additional training data for low-resource Myanmar-English language pairs. Moreover, to address training data sparsity in machine translation tasks, comparable corpora have been shown to be a useful source for creating parallel data that can be used to supplement parallel corpus and can help to improve MT quality. For example, Afli et al. [2016] extracted English-French parallel sentences from a multimodal comparable corpus built from the Euronews[2] and TED[3] websites. Myanmar is a low-resource language and thus the collected English-Myanmar comparable corpora may contain only a small amount of parallel sentence pairs. To obtain more parallel sentence pairs, we first expand the comparable corpora and then extract only the reliable parallel sentence pairs from it.

2.4 Evaluation Methods

Progress of MT relies on accessing the quality of a new system to show that the new system can perform better than previous systems. Human evaluation of MT is a very costly activity, especially considering how fast new systems and their intermediate versions are created and tested. For example, one may want to evaluate tens or hundreds of systems a day to find the best model within models created at each epoch of training or to find the best hyper-parameters that lead to better MT models. Therefore, it is importance to find automatic evaluation metrics since performing manual evaluation is not remotely feasible [Papineni et al., 2002]. In the following sections, we will describe two MT evaluation metrics and one metric for word alignment model used in this work.

2.4.1 Bilingual Evaluation Understudy (BLEU)

BLEU score [Lin et al., 2004a,b] is the current widespread standard metric for automatic evaluation of MT. BLEU measures how many words overlap in a given translation when compared to a reference translation, giving higher scores to sequential words. The BLEU score of MT output is calculated by counting the number of n-grams in the MT output, matched with the set of n-grams in reference translations. The highest n-gram order is defined commonly to be four. For each n-gram order, precision is calculated separately, and the precisions are combined via a geometric averaging as follows:

$$p_n = \frac{\sum_{c \in C_n} Count_{clip}(c)}{\sum_{c' \in C'_n} Count\ (c)} \tag{2.6}$$

Here, C_n indicates a set of n-grams, $Count_{clip}$ truncates count of each word, if necessary, not to exceed the largest count observed in any single reference for that word.

[2] https://www.euronews.com/
[3] https://www.ted.com/

As a result, BLEU score of MT output is calculated as follows:

$$BLEU = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.7}$$

where w_n denotes positive weights summing to one. BLEU scores range from 0-1, the higher the score, the more the translation correlates to a human translation. We express BLEU by multiplying it by 100 in this work.

2.4.2 Translation Edit Rate (TER)

TER [Snover et al., 2006] is a method used by MT specialists to measure the amount of editing that a translator would have to perform to change a translation so it exactly matches a reference translation. TER is defined as the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references. Since it is concerned with the minimum number of edits needed to modify the hypothesis, we only measure the number of edits to the closest reference.

$$TER = \frac{\#\ of\ edits}{average\ \#\ of\ reference\ words} \tag{2.8}$$

TER scores also range from 0-1. However unlike the other scores, with TER a higher score is a sign of more post-editing effort and so the lower score is better, as this indicates less post-editing is required. In this work, we also express TER by multiplying it by 100.

2.4.3 Alignment Error Rate (AER)

AER [Och et al., 2003] is an inadequate measure of word alignment quality. It is commonly used metric for accessing sentence alignments. It combines precision and recall metrics together such that a perfect alignment must have all the sure alignments and may have some possible alignments. Therefore, AER requires a gold standard manually annotated set of "Sure" links and "Possible" links (referred to as *S* and *P*). "Sure" links and "Possible" links are used for measuring Recall and Precision respectively.

Given a hypothesized alignment consisting of the link set A, the three measures: Precision, Recall and AER, are defined:

$$Precision\ (A, P) = \frac{|P \cap A|}{|A|} \tag{2.9}$$

$$Recall\ (A, S) = \frac{|S \cap A|}{|S|} \tag{2.10}$$

$$AER\ (A, P, S) = 1 - \frac{|P \cap A| + |S \cap A|}{|S| + |A|} \tag{2.11}$$

Chapter 3

Optimizing Myanmar Word Segmentation for Translation Performance

3.1 Introduction

For East Asian languages without distinct word boundaries, e.g., Myanmar, Japanese and Chinese, word segmentation is widely applied as pre-processing step in many high-level natural language processing (NLP) tasks (machine translation, part-of-speech tagging, name entity recognition, etc.). Myanmar word segmentation is a difficult task due to the fact there is no unifying segmentation standard.

Myanmar word segmentation is a necessary step in Myanmar-English MT system and its performance has impact on the results of MT. As a common solution in Myanmar-English translation, we segment Myanmar text using an off-the-shelf segmentation tool which is trained on a manually segmented corpus. However, it is not sure that the domain of the segmented corpus exactly matches with the MT task at hand. Consequently, the disambiguation ability of the Myanmar word segmentation will drop, and the performance of the MT system will be influence.

Adapting or optimizing word segmentation has been shown to be helpful for improving translation tasks [Zhao et al., 2013]. If there are many errors in the word segmentation step, a high accuracy translation of the MT system may not be as expected. Existing Myanmar word segmentation tools produce massive out-of-vocabulary (OOV) words or rare words in MT tasks. In order to handle massive OOV words problem, we specifically propose an unsupervised Myanmar word segmentation approach based on the *NFKC*[4] normalization and byte pair encoding (BPE) [Sennrich et al., 2016b] mechanisms. The proposed segmentation approach can learn itself to easily adapt the current MT domain and significantly reduce the OOV rate.

The proposed Myanmar word segmentation model doesn't require any manual work and resources. It enables to learn only on the current MT training data and segment the text that fits with the current MT domain. The proposed segmentation model is trained as follows:

- First, the input Myanmar texts are treated as the raw streams of Unicode characters and normalizes them into canonical forms;
- Then, byte pair encoding (BPE) is applied on the normalized corpus to construct the appropriate vocabulary.

Each step of the proposed segmentation model is described in detail in Section 3.3.

[4] https://en.wikipedia.org/wiki/Unicode equivalence.

The main contribution of this work is that we indicate the feasibility of improving performance on the Myanmar-English machine translations by developing a simple but effective unsupervised Myanmar word segmentation model. Experiments on machine translation systems evaluate the effectiveness of the proposed Myanmar word segmentation model on improving current MT systems.

The remainder of this work is organized as follows: Section 3.2 reviews the prior research on Myanmar word segmentation and their limitations, Section 3.3 introduces the proposed architecture of Myanmar word segmentation, Section 3.4 explains the dataset, experimental setup, and discusses the results of the experiments, and section 3.5 concludes our work.

3.2 Related Work

Previous works on Myanmar word segmentation are described in this section. There are a lot of research works that has been done on the Myanmar word segmentation problem and many approaches have been proposed. The former research work is implemented with a simple forward maximum matching approach and used extra rules with cases [Pa and Thein, 2008]. While there are few linguistically annotated resources for Myanmar; thus, only dictionary- and rule-based approaches are feasible. Then, Thu et al., [2014] proposed a word segmentation approach integrating dictionaries and unsupervised Bayesian models. However, in their report, the best segmentation performance was still achieved by forward maximum matching. Later, the statistical, machine learning and hybrid approaches are proposed for Myanmar word segmentation problem [Ding et al., 2016; Phyu and Hashimoto, 2017; Oo and Soe, 2019].

In the dictionary-based Myanmar word segmentation model, only the words that are added in a pre-defined dictionary can be identified and thus the performance of the model depends mainly to a large degree upon the coverage of the dictionary. For solving the out-of-vocabulary (OOV) words problem, increasing the size of the dictionary is not a perfect solution because the new words appear constantly. On the other hand, although the statistical-based models can somehow solve the problem of OOV words in segmentation process by exploiting probabilistic or cost-based scoring mechanisms, these models also suffer from some drawbacks. The main challenges are that these models require huge amounts of training data with an amount of the processing time. Moreover, these models have the difficulty in incorporating the linguistic knowledge effectively into the word segmentation process [Teahan et al., 2000].

Myanmar is a resource-poor languages and thus there are only corpus-based, dictionary-based, rule-based, and statistical-based word segmentation tools are freely available for being used as a temporary solution for Myanmar texts segmentation. Although, the current segmentation models can support to achieve better results for some language processing tasks, such as part of speech (POS) tagging, word sense disambiguation (WSD), text categorization, information retrieval (IR), text summarization, etc. However, these models may probably produce massive rare-words or OOV words in the current MT systems. Especially, the Myanmar segmentation errors would cause translation mistakes directly in English-to-Myanmar MT translation direction. Although it is not a very serious problem in Myanmar-to-English translation direction, weak Myanmar word segmentation tools can lead SMT system to produce unknown source words as target translated words in the target translations. This is

because they cannot find the corresponding target translation for OOV source words in the phrase table. In NMT systems, the same problem also occurs.

Figure 3.1 illustrates the examples of translation errors generated by the current English-to-Myanmar SMT and NMT systems while using the currently available Myanmar word segmentation model. There are three kinds of translation errors. They are: (i) missing words or phrases in the target translations, (ii) translating source (English) words into wrong target (Myanmar) words, and (iii) generating both source English words and their correspondence target Myanmar words together in the translations.

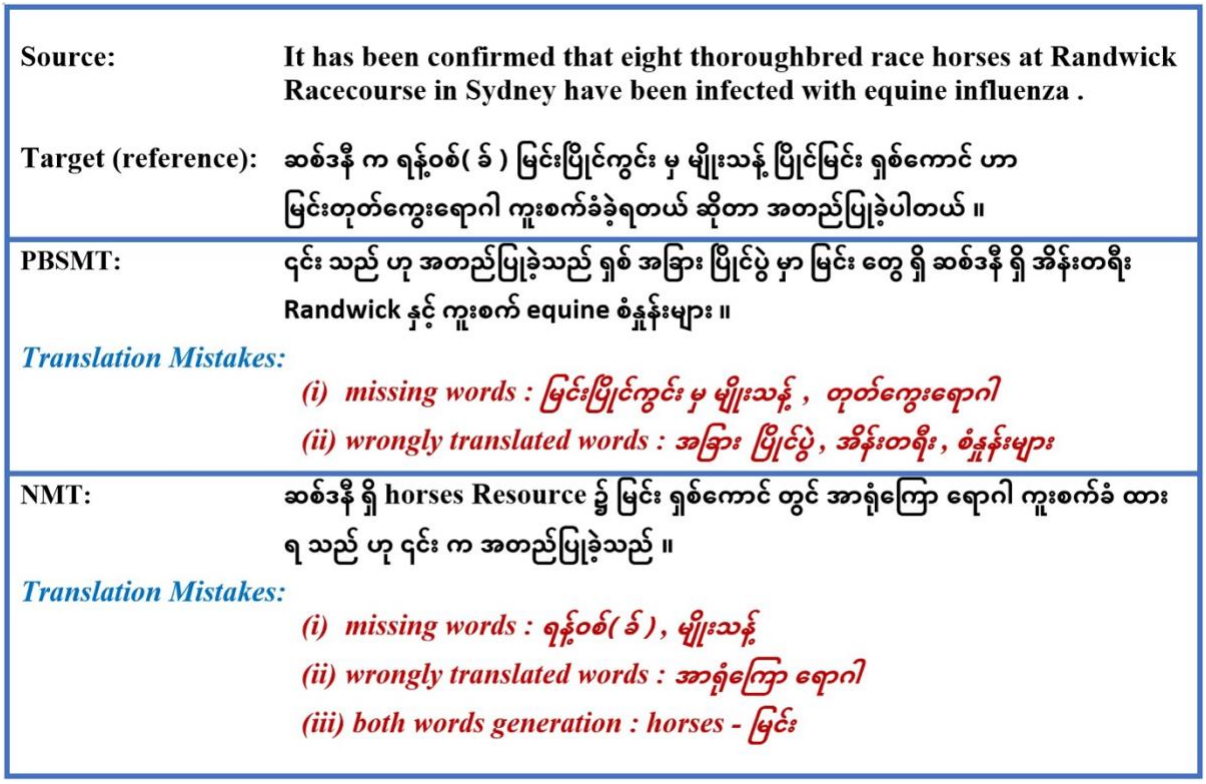

Figure 3.1: Translation errors of both statistical and neural English-to-Myanmar MT systems due to the Myanmar word segmentation weakness.

As in Figure 3.2, in Myanmar-to-English translation direction, even in the translation of a short sentence, SMT copies an unknown Myanmar source word (in red) and pastes it in the target sentence translation. On the other hand, NMT system completely misses to translate this source word in its translation. Actually, his Myanmar source word (in red) should be translated as “fifteen” in the target English translation. This Myanmar word consists of the two sub-words (one word in green and another word in blue). The word (in green) means “fifteen” in English and the other word (in blue) is a numerical classifier and thus it has no special meaning in English. This kind of numerical classifier is used only in Myanmar language. It usually follows a number to show what type of thing that the number is referred to. These errors are evolved because the current Myanmar word segmenter can only segment words based on their trained nature, corpus and dictionary. Therefore, Myanmar word segmentation using the currently

available tool may not be fit with the training corpus intended to use in MT tasks. In the above examples, Myanmar texts are segmented using off-the-shelf segmentation tool, UCSYNLP word segmenter[5], that is implemented by a combined model of bigram with word juncture and it works by longest matching and bigram methods trained on a pre-segmented corpus of manually collected 50,000 words from Myanmar text books, newspapers, and journals [Pa and Thein, 2008].

The proposed Myanmar word segmentation approach does not require any linguistic resources, preprocessing, and manual works. The only requirement is to convert the sentence into Unicode encoding. However, for this requirement, there are many Myanmar Unicode converters that are currently freely available online and offline. The proposed segmentation model enables to learn on current MT data and thus it can generate the most suitable word segmentation results. The effectiveness of the proposed word segmenter is evaluated by conducting the MT experiments. For evaluation metric, BLEU score is used.

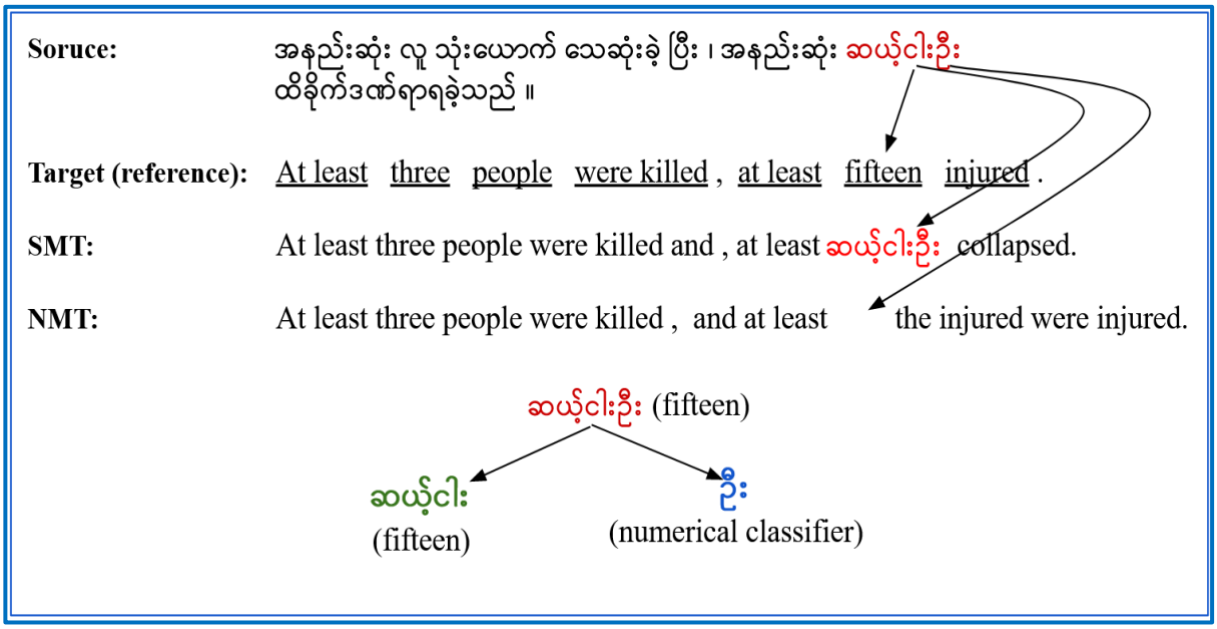

Figure 3.2: Translation errors of both statistical and neural Myanmar-to-English MT systems due to the Myanmar word seg- mentation weakness.

3.3 Unsupervised Myanmar Word Segmentation

The proposed segmentation model consists of three main components: a normalizer, a trainer, and a tokenizer. We treated the input sentences as the raw Unicode character streams, including the space as a use character. Figure 3.3 depicts an overall architecture of the proposed word segmentation model.

[5] http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/wsandpos.html

The first module, the normalizer (indicated by the first blue box) used the Unicode *NFKC* normalization to normalize semantically equivalent Unicode characters into the canonical forms. *NFKC* is the Unicode standard normalization form that has been widely used in various downstream NLP tasks recently because of its better reproducibility and its strong support on the Unicode standard. The second module, the trainer (indicated by the second blue box) trains the segmentation model utilizing the byte-pair-encoding (BPE) mechanism [Sennrich et al., 2016b] from the normalized corpus to build up a word vocabulary based on the sub-word components. The trained segmentation model helps learning a vocabulary and provides a good compression rate of the text. The final tokenizer module (indicated by the dashed box) internally executes the normalizer to normalize the input sentences and tokenizes them into a sub-word sequence with the word segmentation model trained by the trainer.

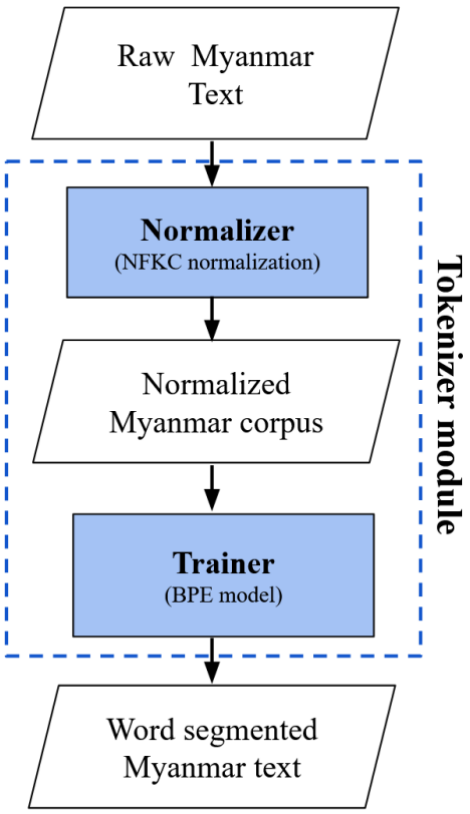

Figure 3.3: The proposed Myanmar word segmenter.

3.4 Experiments

This section describes the datasets, baseline MT systems that we have used in this work. Then, the detail analysis of experimental results is discussed.

3.4.1 Datasets

We collected around 224 thousand manually created English-Myanmar parallel sentences from text books, Myanmar local news, and the ALT Corpus [Riza et al., 2016] and used them for training data. The development and test sets used in this work are only from the ALT corpus. Data statistics are shown on Table 3.1.

Table 3.1: Statistics of parallel datasets.

Type	Data Source	Total Sentences
Train	Local News and Textbooks ALT	204,535 18,082
Dev	ALT	1,000
Test	ALT	1,017

3.4.2 Baseline MT systems

The effectiveness of the proposed Myanmar word segmentation model is evaluated by conducting machine translation experiments. We used the following MT systems as baseline systems in this work.

Statistical Machine Translation:

We used Moses toolkit [Koehn et al., 2007] for training the phrase-based SMT (PBSMT) system. GIZA++ [Och and Ney, 2003] tool is used to implement the word alignment process. We used grow-diag-final and msd-bidirectional-fe heuristic for phrases extraction and lexicalized word reordering. For tuning the PBSMT model, we used the default parameters of Moses. Moreover, the 5-gram language models were trained on Myanmar and English monolingual texts with Kneser-Ney smoothing using KenLM [Heafield et al., 2013] tool.

Neural Machine Translation:

For NMT, we trained a Transformer-based model with PyTorch version of the OpenNMT toolkit, an open-source (MIT) NMT framework [Klein et al., 2018]. The Transformer experiments were run on NVIDIA Tesla P100 GPU by using the following parameters listed in Table 3.2.

3.4.3 Experimental Results

To evaluate the proposed Myanmar word segmentation model, we conducted the experiments using SMT and NMTsystems: PBSMT and Transformer. We use BLEU (Bilingual Evaluation UnderStudy) score as the evaluation metric. BLEU scores are computed using the multi-bleu script from Moses toolkit.

Table 3.2: Parameters for training Transformer models.

-layers 6 **-rnn_size** 512 **-word_vec_size** 512 **-transformer_ff** 2,048 **-heads** 8 **-encoder_type** transformer **-decoder_type** transformer **-position_encoding** true **-train_steps** 200,000 **-max_generator_batches** 2 **-dropout** 0.1 **-batch_size** 4,096 **-batch_type** tokens **-normalization** tokens **-accum_count** 2 **-optim** adam **-adam_beta2** 0.998 **-decay_method** noam **-warmup_steps** 8,000 **-learning_rate** 2 **-max_grad_norm** 0 **-param_init** 0 **-param_init_glorot** true **-label_smoothing** 0.1 **-valid_steps** 1,000 **-save_checkpoint_steps** 1,000 **-world_size** 1 **-gpu_rank** 0

In our experiments, Moses tokenizer and truecaser are used to tokenize the English sentences. For Myanmar word segmentation, UCSYNLP word segmenter is applied as a baseline word segmentation model. As explained in Subsection 3.3, the proposed segmentation model consists of three main modules: a normalizer, a trainer, and a tokenizer. For the normalizer and trainer modules, the same modules of Unicode *NFKC* Normalizer and BPE Trainer provided by SentencePiece [Kudo and Richardson, 2018] are used. In the trainer process, a vocabulary size of 32,000 BPE sub-words is used. Our tokenizer module internally executes the normal- izer to normalize the input Unicode character streams and tokenizes them into the word sequences with the segmentation model trained by the trainer.

Table 3.3: BLEU scores of English-to-Myanmar translation systems on two segmentation models (baseline and ours).

	UCSYNLP Segmenter	**Our Segmenter**
SMT	4.15	7.63
NMT	5.25	8.11

Table 3.4: BLEU scores of Myanmar-to-English translation systems on two segmentation models (baseline and ours).

	UCSYNLP Segmenter	**Our Segmenter**
SMT	9.41	9.19
NMT	10.24	11.59

Based on the proposed Myanmar word segmentation model, the performance of MT systems are quite different. The results reported in Table 3.3 and Table 3.4, showing that our proposed unsupervised Myanmar segmentation model can support both SMT and NMT systems to significantly outperform the previous baselines. According to the experimental results, using the proposed segmentation model has large gains on both SMT and NMT systems in both translation directions. In the English-to-Myanmar translation direction, with our

proposed segmentation model, SMT and NMT achieved a BLEU score of 7.63 and 8.11, respectively, which outperforms the previous best result by +3.48 and +2.86 points. In Myanmar-to-English translation direction, NMT system still outperforms the baseline score by +1.35 BLEU points. However, the score of SMT system is slightly decreased from 9.41 to 9.19. The decrease in BLEU score is happened because of the names and numbers that are not specifically cared during the segmentation process. Some of the rare names and numbers in the sentences are segmented into two or three sub-words, and is thus led to a little weakness only in the word alignment procedure of Myanmar-to-English PBSMT. Table 3.5 reports the number of OOV words in Dev and Test of ALT dataset. Our proposed segmentation significantly reduces the number of OOV words than the baseline model.

Table 3.5: Number of OOV words in Dev and Test data of ALT dataset on two segmentation models (baseline and ours).

	ALT Dev		**ALT Test**	
	Unique Words	OOV words	Unique Words	OOV words
UCSYNLP Segmenter	7683	3399	8031	3606
Our Segmenter	5990	27	6204	29

To examine the OOV words problem, we exploited a copy mechanism in all experiments. The copy mechanism first tries to substitute the OOV words with the target words that have maximum attention weight according to their source words [Luong et al., 2015]. When the target words are not found, it copies the source words to the position of the not-found target words [Gu et al., 2016]. A detailed study of the results in English-Myanmar bi-directional translations reported that the number of OOV words decreased significantly with our proposed segmentation model. Some example sentences produced by the English-to-Myanmar MT systems with the baseline segmentation model(UCSYNLP segmenter) and with our word segmenter are shown in Figure 3.4. Both SMT and NMT systems with our Myanmar word segmentation model could handle the problem of OOV words than the outputs with baseline segmenter. In figure, the blue colored parts of the sentences are the correct translation parts in Myanmar language. Overall, NMT systems lead to better translation accuracy and fluency than SMT system. After obtaining the experimental results, in all cases, we concluded that NMT systems with the proposed Myanmar word segmentation model is the best that can produce more accurate and fluent translations (cf. Figure 3.4).

English → Myanmar	
Source	It has been confirmed that eight thoroughbred race horses at Randwick Racecourse in Sydney have been infected with equine influenza .
SMT (UCSYNLP)	၎င်း သည် ဟု အတည်ပြုခဲ့သည် ရှစ် အခြား ပြိုင်ပွဲ မှာ မြင်း တွေ ရှိ ဆစ်ဒနီ ရှိ အိန်းတရီး Randwick နှင့် ကူးစက် equine စံနှုန်းများ ။
SMT (Ours)	(၈) ကောင် ကို အတည်ပြု ပြီး ၎င်း သည် တုပ်ကွေး ရောဂါ ပြိုင်ပွဲ မှာ မြင်း တွေ က ဆစ်ဒနီ မြို့ တွင် Randwick ဖြင့် ကူးစက် equine ခံ ခဲ့ ရတယ် ။
NMT (UCSYNLP)	ဆစ်ဒနီ ရှိ horses Racecourse ၌ မြင်း ၈ ကောင် တွင် အာရုံကြော ရောဂါ ကူးစက် ခံထား ရ သည် ဟု ၄င်း က အတည်ပြုခဲ့သည် ။
NMT (Ours)	ထို ဆစ် ဒ် နီ တွင် ဆစ်ဒနီ မြို့ ၌ ပြိုင်မြင်း ၈ စီး ကို မြင်းတုတ်ကွေး ကူးစက် ခံခဲ့ရ ကြောင်း အတည်ပြု ခဲ့ ပြီး ပြီ ဖြစ်သည် ။
Reference	ဆစ်ဒနီ က ရန်ဝစ်(ခ်) မြင်းပြိုင်ကွင်း မှ မျိုးသန့် ပြိုင်မြင်း ရှစ်ကောင် ဟာ မြင်းတုတ်ကွေးရောဂါ ကူးစက်ခံခဲ့ရတယ် ဆိုတာ အတည်ပြုခဲ့ပါတယ်။
Source	After the explosion, they commented that workers were complaining the ventilation of the department is poor.
SMT (UCSYNLP)	အဆိုပါ ပေါက်ကွဲမှု ပြီးနောက် ၊ သူတို့ ညည်း နေ ကြ သည် ဟု မှတ်ချက် ပေးခဲ့သည် ဟု ၊ အလုပ်သမားများ သည် လေ ဟာ ဆင်းရဲ နွမ်းပါး သူများ ၏ အဆိုပါ ဌာန ကို ပြောခဲ့သည် ။
SMT (Ours)	အဆိုပါ ပေါက်ကွဲမှု ပြီးနောက် ၊ သူတို့ သည် ဟု ညည်း နေ ကြ သည် ဟု မှတ်ချက်ချခဲ့သည် အလုပ်သမား ဦးစီးဌာန သည် လေဝင် လေ ထွက် ညံ့ဖျင်း က ပြောခဲ့သည် ။
NMT (UCSYNLP)	ပေါက်ကွဲမှု ပြီးနောက် ၊ သူတို့ က ဌာန ၏ workers ကို အလင်းရောင် လေဝင် လေ ထွက် လို့ မှတ်ချက် ပေးခဲ့သည် ။
NMT (Ours)	ပေါက်ကွဲမှု ပြီးနောက် ၊ သူတို့ က ဌာန ၏ လေ ဝင် လေ ထွက် ကောင်းမွန် မှု မရှိ ကြောင်း ဝေဖန် ပြောဆိုခဲ့သည် ။
Reference	ပေါက်ကွဲမှု ဖြစ်ပြီးနောက်၊ အလုပ်သမားများ က ထို ဌာန၏ လေဝင်လေထွက်ခြင်း မကောင်းပါ ဟု ညည်းညူခဲ့သည်ဟု ထင်မြင်ချက် ပေးခဲ့သည်။
Source	None of the killings had been investigated satisfactorily , Colville said .
SMT (UCSYNLP)	မဟုတ်တာ အဆိုပါ သတ်ဖြတ်မှု များ သည် ၊ ကျေ ကျေနပ် စုံစမ်း Colville က ပြောခဲ့သည် ။
SMT (Ours)	မည်သူမျှ မ သိ ကွာ အရ သတ်ဖြတ် ခြင်း ခံ ခဲ့ ရ ပြီး ၊ ကို စုံစမ်း စစ်ဆေး သည့် အခါ ကျေနပ် Colville က ပြောခဲ့သည် ။
NMT (UCSYNLP)	လူသတ်မှု ၏ အဖြေ ကို killings ခြင်း တစ် ခု မှ အပေါ်တွင် စုပ် ယူ ခြင်း မ ရှိခဲ့သည် ၊ ဟု အယ် Colville က ပြောသည် ။
NMT (Ours)	လူသတ်မှု တွေ ထဲက ဘယ် သူကို မျှ ကျေ ကျေနပ် စုံစမ်း သေး ဟု ၊ လီ ဘ မန် က ပြောခဲ့သည် ။
Reference	ဘယ် သတ်ဖြတ်မှုများ ကိုမျှ စိတ်ကျေနပ်ဖွယ် စုံစမ်းစစ်ဆေးခြင်း မပြုလုပ်ခဲ့ပါ ဟု ၊ ကော်မစ်လီ က ပြောကြားခဲ့သည် ။

Figure3.4: Example of translations in English-to-Myanmar direction using SMT and NMT. The translation performance of both MT systems improved with our segmentation model compared to the baseline UCSYNLP segmenter. NMT with our segmentation approach generates more accurate and fluent translation outputs.

3.5 Summary

This work is motivated by our expectation of improving the translation performance of the current English-Myanmar MT systems while using the available limited resources. To achieve this goal, we present our main contributions in Myanmar word segmentation. The proposed word segmentation approach is trained using the Unicode *NFKC* normalization and the byte-pair-encoding (BPE) mechanism. The proposed segmentation approach is aimed to improve the performance of the translation systems by reducing the out-of-vocabulary (OOV) rate. We evaluated the performance of proposed segmentation approach by conducting both SMT and NMT experiments.

Unlike the previously popular Myanmar word segmentation models that make use of manually prepared resources such as large-scale training data, dictionaries, etc, our proposed model does not need any manual work and also any knowledge about Myanmar language. The

only requirement in this model is to convert Myanmar text written in other fonts into Unicode fonts. Conversion can be done with the use of freely available tools. Using the proposed Myanmar word segmentation model on the preprocessing step of NMT systems, the performance of their translations improved quite a lot.

Chapter 4

Automatic Post-editing for Machine Translation

4.1 Introduction

Although recent advances have reported that neural machine translation (NMT) systems achieve near human-level performance on several language pairs, their translations are not always perfectly correct. Human experts are required for correcting the systematic errors in machine-translated texts. Automatic Post Editing system aims to automatically fix the translation errors by learning from human post-edited samples. The phrase-based statistical machine translation (PBSMT) models are used in earlier APE research to train the APE system as a monolingual re-writing task without considering the source sentence [Simard et al., 2007a,b]. However, these models are only applicable to fix the errors in the translations of rule-based MT systems. While applying the PBSMT model both for first-stage MT and the second stage APE, there are no or only modest improvements without additional source context modelling and thresholding [Béchara et al., 2011]. Most recent APE researchers adopt a dual-source (or multi-source) sequence-to-sequence structure that extends the Transformer [Vaswani et al., 2017] in a supervised learning setting [Chatterjee et al., 2018; 2019].

Training an APE system general requires a training set comprising the triplets (source-text, MT-output, human post-edit), denoted as ⟨*src, mt, pe*⟩, respectively. APE system simultaneously takes the source sentence (*src*) and its corresponding MT output (*mt*) as inputs and use the associated human post-edited sentence (*pe*) as the target. For training the deep and complex APE models, the quantity of available APE data is still insufficient due to the high production cost of the target data (*pe*). Moreover, the current APE systems have failed to show any notable improvement in the refinement of NMT translations when training on similar-sized human post-edited data [Chatterjee et al., 2018; 2019; Ive et al., 2020].

APE triplets are publicly available only for very few language pairs such as English-German and English-Chinese[6]. Most language pairs absent of APE data and thus hinder the applicability of the APE task. In this work, we investigate an alternative solution to conduct the APE research without having access to the APE triplets.

The proposed APE system is trained using only monolingual and parallel MT corpus. It doesn't require any human-edited APE triplets. The MT output (*mt*) and its original source sentence (*src*) are taken as the inputs into the APE system, and the high-quality target sentence (*post-edited mt*) is generated as output by performing a series of the following three steps:

1. Extract word alignment information of the input pair (*src* and *mt*) such as alignments and unaligned source and target words for the input pair using the proposed word aligner,

[6] https://statmt.org/wmt21/ape-task.html

2. Enrich the *mt* by removing the errors (unaligned target words) from it and adding missing information (word-to-word translation of unaligned source words to target words using bilingual dictionaries), and
3. Denoise the enriched version of *mt* (output from Step 2) to transform it into a high-quality target sentence with the proposed denoiser.

The proposed word aligner exploits LaBSE [Feng et al., 2020] based word embeddings to align the similar source and target word. The aligner records both alignments and unaligned words. In the sentence enrichment module, we create and use two types of bilingual dictionaries by leveraging monolingual and parallel corpus. In denoising module, Transformer [Vaswani et al., 2017] based models are trained on target monolingual data and used them to denoise the errors in the input sentence. The following are the main contributions of this work are:

- A new unsupervised word aligner is developed by leveraging LaBSE based word embeddings. The proposed model performs better than the previous models even in the absence of explicit training on parallel data.
- A simple and effective enrichment module is designed to enrich raw MT texts by exploiting the bilingual dictionaries created from existing monolingual and parallel corpus.
- A robust denoiser is proposed and used as a final postprocessor in the APE systems. The denoiser can transform the input sentence into a qualified output by handling multi-aligned words and local reordering.
- We prove that sub-word level embeddings perform poorly in word alignment task.
- We practically show that an APE system built by combining the three modules: word aligner, sentence enrichment module, and denoiser; is effective to correct the systematic errors in MT texts. The proposed approach can be the best one to develop the reliable APE system in a low-resource setting where APE triplets are unavailable.

For most language pairs, where human-edited APE triplets are unavailable, the proposed APE system can be effectively used as a post-processor to the raw MT text. In this work, we provide the detail analysis to show the better understanding of leveraging the pre-trained model in the APE task to produce bilingual word embeddings and using these embeddings for extracting word alignment information and enriching the translated sentences. We additionally demonstrate that the denoising autoencoder can be applied not only as a language model in various natural language processing (NLP) tasks but also as a monolingual sentence rewriter in the APE system. According to the results of our experiments, the proposed unsupervised word aligner outperforms the existing state-of-the-art alignment models on the word alignments extraction task. Moreover, the proposed APE framework can post-edit the raw translated texts generated by existing English-Myanmar MT systems to meet the agreed level quality. We believe that this work can provide the optimal research direction in the study of APE for most of the language pairs where APE data are unavailable.

4.2 Denoising-based Automatic Post-editing System

The proposed denoising-based APE system (DbAPE) is a pipeline consisting of three major modules. The first module depicted by Figure 4.1 (a) take the source and target sentence pair (i.e., *src* and *mt* pair) and then extracts their alignments. This module also records the unaligned source and target words. Figure 4.1 (b) depicts the second module for sentence enrichment. The enrichment module minimizes the semantic gap between src and *mt* by removing the typical errors (i.e., unaligned words) in *mt* and enriching *mt* with the missing source-side information (i.e., translate unaligned source *src* words into target *mt* words and append them to *mt)*. The final module called target sentence denoising, is depicted by Figure 4.1 (c). This module takes the output of the previous module (i.e., *enriched-mt*) as input and remove the noises and correct the words and phrase order.

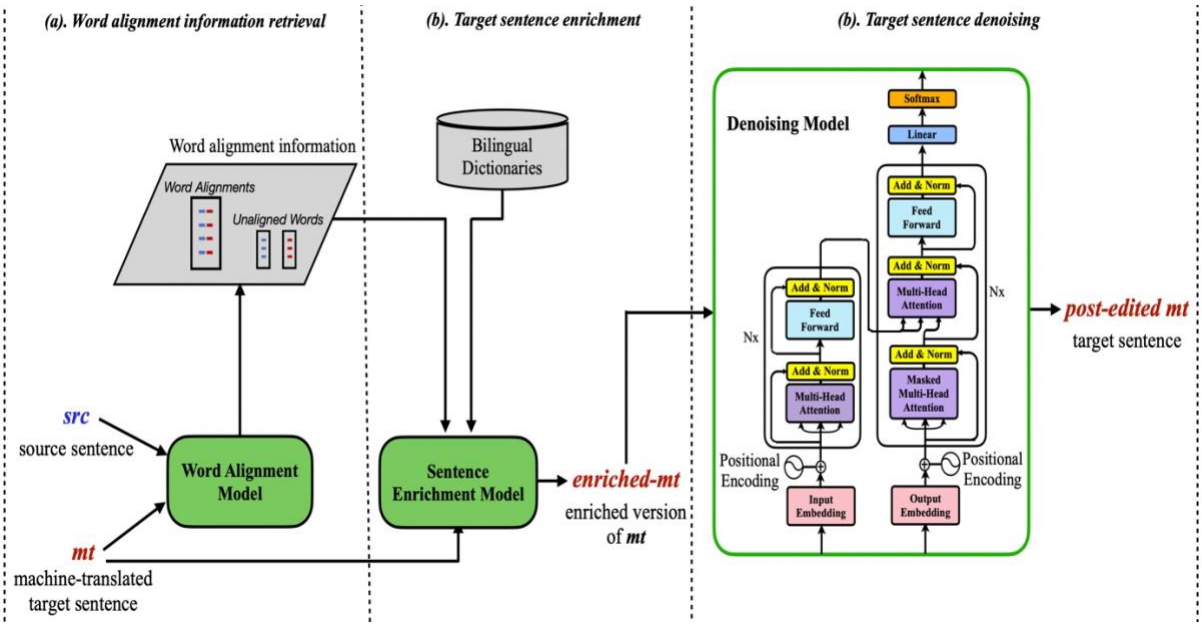

Figure 4.1 : Overall architecture of denoising-based automatic post editing (DbAPE) system

4.2.1 Word Alignment Information Retrieval

Given a pair of source sentence $x = (x_1, x_2, \dots, x_n)$ of length n and its corresponding parallel target sentence $y = (y_1, y_2, \dots, y_m)$ of length m, the task of word aligner A is to find a set of pairs of source and target words which are semantically similar to each other within the context of the sentence.

$$A = \{< x_i, y_j >: \ x_i \in x, \ y_j \in y\} \tag{4.1}$$

4.2.1.1 Extracting Alignments from Embeddings

The pre-trained word embedding models such as RoBERTa [Liu et al., 2019] and BERT [Devlin et al., 2019] represent words using continuous vectors calculated in context and have achieved impressive performance in a various downstream NLP tasks. The sentence embedding models such as language-agnostic BERT, called LaBSE [Feng et al., 2020], have multilingually trained by adapting multilingual BERT (mBERT) [Devlin et al., 2019] to produce language-agnostic cross-lingual sentence embeddings for 109 languages. LaBSE model has been achieved the state-of-the-art performance on the parallel text (bi-text) mining task. Although LaBSE is originally developed for bi-text mining to find the translation pairs in multiple languages, this work uses it for the word alignment extraction task. Based on the LaBSE based word embeddings, we identify and extract semantically all similar source and target word pairs in a given parallel sentence pair.

Algorithm 1: Word Alignment Information Retrieval

$src = [src_1, src_2, \dots, src_n]$
$mt = [mt_1, mt_2, \dots, mt_m]$
$A_{src-mt} = [\,]$
$A_{src} = [\,]$
$A_{mt} = [\,]$
$U_{src} = [\,]$
$U_{mt} = [\,]$
for $i = 1$ to m:
    for $j = 1$ to n:
        $k = 0$
        $highest\_score = 0$
        $score = cosine\_sim(\text{LaBSE}(mt_i)\ , \text{LaBSE}(\ src_j))$
        if $score > highest\_score$:
            $highest\_score = score$
            $k = j$
    if $highest\_score \geq t$:
        append $(src_k\ , mt_i)$ to A_{src-mt}
        append src_k to A_{src}
        append mt_i to A_{mt}
for $i = 1$ to m:
    if $mt_i \notin A_{mt}$:
        append mt_i to U_{mt}
for $j = 1$ to n:
    if $src_j \notin A_{src}$:
        append src_j to U_{src}

While the prior word alignment models have relied on large-scale parallel data to obtain the correct alignments, here we propose a simpler and more effective approach which is practically suitable for low-resource languages where the large-scale training data is unavailable. In this

step, a robust word alignment model is proposed to align similar source and target words based on their LaBSE word embeddings. This word alignment extraction task is considered as a semantic search task.

In the reminder of the section, we denote the list of word alignments, the list of aligned source words, and the list of aligned target words by A_{src-mt} , A_{src} , and A_{mt}, respectively. The lists of unaligned source words and target words are also denoted by U_{src} and U_{mt}. The detail procedure of the word alignment model is described in Algorithm 1, where t is a user-defined word pair similarity threshold. As the cosine similarity score between the word vectors falls in the range of 0 to 1, the value of threshold *t* is set to 0.5, at the halfway mark.

According to Algorithm 1, the task of word alignment information retrieval proceeds as follows. When a pair of source sentence (*src*) and its corresponding MT output (*mt*) is given, the most similar *src* word is extracted for each *mt* word. The similarity score of each word pair is computed by the cosine similarity function based on their LaBSE word embeddings. Among the extracted source-target word pairs, the pairs with the score higher than the pre-defined threshold are added to the list of final word-aligned pairs A_{src-mt}. Meanwhile, the unaligned source words U_{src} and unaligned target words U_{mt}, in *src* and *mt*, are recorded respectively.

4.2.2 Target Sentence Enrichment

The raw machine-translated texts are enriched in this section by removing their systematic errors and adding missing source-side information to the texts. The errors and missing information are obtained from the word alignment information retrieval task. Bilingual dictionaries are used to transform missing source-side information into target language words.

Bilingual dictionaries (cf. Figure 4.1) are built from the monolingual and parallel corpus. There are two types of bilingual dictionaries: a monolingual corpus-based dictionary (*MD*) and a parallel corpus-based dictionary (*PD*).

A bilingual *MD* is created from the source and target monolingual texts. The procedure to create *MD* is as follows:

- First, the source and target vocab files which contain the list of source and target words are created from source and target text, respectively,
- Then, these two vocab files are feed as input into the proposed word aligner (Algorithm 1), and store all extracted source-target aligned pairs in *MD*.

A bilingual *PD* is created from the existing parallel corpus. For each pair of source sentence x and target sentence y in the parallel corpus, the word alignments are extracted as follows:

- Firstly, the forward-aligned $A_{forward}$ and backward-aligned $A_{backward}$ word pairs are created by running the proposed word aligner (described in Algorithm 1) in source-to-target and target-to-source directions, respectively, as follows:

$$A_{forward} = \{< x_i, y_j >: \ x_i \in x, \ y_j \in y\} \tag{4.2}$$

$$A_{backward} = \{< x_i, y_j >: \ x_i \in y, \ y_j \in x\} \quad (4.3)$$

- Then, the common aligned word-pairs $< x_i, y_j >$ from both lists $A_{forward}$ and $A_{backward}$ are extracted and stored them in *PD*, i.e.,

$$PD = \ A_{forward} \cap A_{backward} \quad (4.4)$$

Since the bilingual direction, *PD* is built with the guidance of the parallel aligned sentence pair in the supervised setting, it should be more accurate than *MD* built in the unsupervised setting and we confirm this hypothesis from our experiments.

While having the raw translated sentence (*mt*), the list of unaligned source words (U_{src}), the list of unaligned target words (U_{mt}), and the two bilingual dictionaries (*PD* and *MD*), the model first deletes the unaligned target words in *mt* according to U_{mt}. Then, the most similar target words of U_{src} are extracted from the bilingual dictionaries and append them to the end of *mt*. The final output of this module is an enriched-version of *mt* (*enriched-mt*).

The process of extracting the most similar target word tor each unaligned source word w_{src} in the list U_{src} , is as follow:

- If w_{src} is in the source-side words of bilingual PD, its aligned target-side word is from *PD* is extracted.
- Else if w_{src} is not in the bilingual *PD* but it is in *MD*, its aligned target word from *MD* is extracted.
- Else, the most similar source word of w_{src} is find in *PD* first and its aligned target word from *PD* is extracted.

If the source word is aligned to more than one target words in the bilingual dictionary, only the target word that has the highest similarity is extracted. Figure 4.2 illustrates this approach.

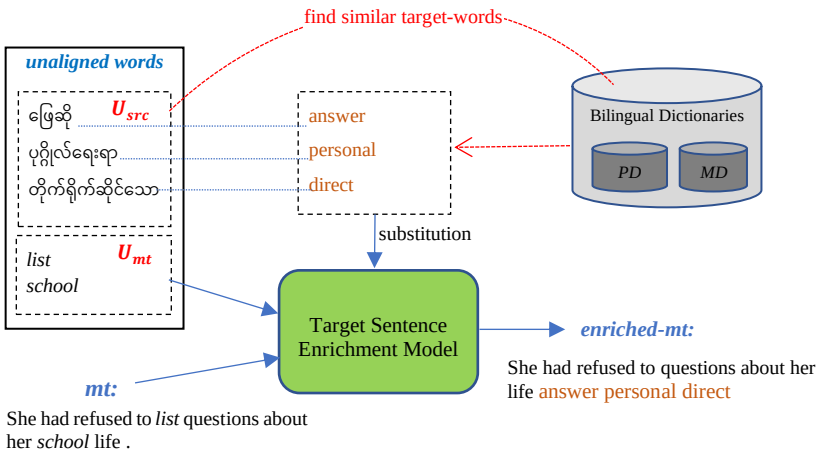

Figure 4.2: Example of Target Sentence Enrichment Task.

4.2.3 Target Sentence Denoising

Although the previous module has removed the mis-translated or extra words and added missing source-side information in target words, the enriched-version of the MT output (i.e., *enriched-mt*) is still far from being an acceptable translation. It still needs for improving the word and phrase order and performing the grammar correction. Furthermore, in the appended part of the *enriched-mt*, the substitution of unaligned source word to similar target word always outputs a target word for every position. In some cases, one or two of the substituted (appended) words should be remove/denoise to make a fluent sentence. On the other hand, it needs to add extra common words, e.g. prepositions or articles, to be in the correct sentence structure. For example, the word-to-word substitution would be substituted a sequence of Myanmar source words "နှစ်ယောက်စလုံး သူတို့ ကို" with a sequence of the similar target English words "both them to"; however, it must be "both of them" in English language. In this case, the substituted word "to" can be considered as an *insertion noise* that needs to remove from the sentence and the extra word "of" is considered as a *deletion noise* that must be added to the sentence.

A sequence-to-sequence Transformer [Vaswani et al., 2017] model is applied for removing the potential noises in the *enriched-mt*. The model takes a noisy (unstructured) text as input and generates a clean (denoised) text as output; both of which are of the same (target) language. As shown in Figure 4.1 (c), the noisy input, *enriched-mt*, is fed into a designed denoising model to transforms it into a clean target sentence (i.e., *post-edited mt)*. For investing the effectiveness of various denoising approaches on improving the quality of the APE final output, the experiments on the denoising task are conducted with the following two different denoising models.

4.2.3.1 Denoising Autoencoder

Training label sentences would be the target monolingual sentences to train a denoising autoencoder. Given a clean sentence as the target, the noisy input to the denoiser should be ideally the unstructured version of that sentence. For creating the noisy sentences, some artificial noises are injected into the clean sentences. These artificial noises are created based on the noises found in the enriched-version sentences.

First, 20 to 30 percent of out-of-vocabulary (OOV) words are removed from each sentence in the monolingual corpus. Then the deleted words are appended to the end of the sentence. Finally, the following artificial noises are injected:

a) Insertion of random frequent tokens into the sentences helps the model to learn, identify, and remove the extra/redundant words from theses sentences:

   1. For each position i, a probability p_i ~ Uniform (0, 1) is first sampled,
   2. Let p_{ins} be a probability threshold of the insertion. If $p_i < p_{ins}$, a word w from the most frequent target words V_{ins} is sampled and then inserted it before the position i.

   We limit the inserted words by V_{ins} because target insertion occurs mostly with common words, e.g., prepositions or articles. For deciding whether the words should be inserted or not, we threshold the value with p_{ins}.

b) Deletion of tokens from the sentences helps the model learn to predict and add potential words to these sentences for fluency:

   1. For each position i, a probability p_i ~ Uniform (0, 1) is first sampled,
   2. Let p_{del} be a probability threshold of the deletion. If $p_i < p_{del}$, the word at the position i is dropped.

   For deciding whether the words should be deleted or not, we threshold the value with p_{del}.

c) Permutation of tokens with a limited distance in the sentences is applied to stimulate the learned model to modify the word order in a correct target structure:

   1. Let d_{per} be a degree of the permutation. For each position i, an integer $\delta_i \in [0, d_{per}]$ is sampled,
   2. We add δ_i to index i and sort the incremented indices $i + \delta_i$ in an increasing order,
   3. The words are rearranged in the new positions, to which their original indices have moved by Step 2.

The designs and settings of the previous work in [Kim et al., 2018] are adopted for insertion, deletion, and reordering noises. In the denoising module, a vocabulary size of 32,000 words is considered. The words out of this vocabulary are called OOV words.

4.2.3.2 Denoising Rewriter

A Transformer-based target-to-target rewriting model is proposed and trained it to learned and produce a clean sentence from its noisy-version. For training the this model, we build the noisy training data from the target monolingual corpus D_y. First, 20 to 30 percent of the OOV words are removed from a given sentence $y \in D_y$, and then the removed words are appended to the end of y. Next, for building the insertion and deletion noise types, some words (up to three words for the sentences with the sentence-length greater than ten) are randomly dropped/added. Finally, contiguous words are swap randomly with a probability p_{swap} to inject some noises to get the noisy version y'. For this denoiser, the value of the p_{swap} is set to 0.2. We treat the noisy version y' as the input and the original clean sentence y as the output to train the denoiser. For model inference, the enriched-version of MT text (*enriched-mt*) is fed into the trained model and generate the clean target sentence, *post-edited mt*.

4.3 Experiments

In this section, the detail about the evaluation metrics, datasets and model configuration used in this research work are described. Then, the experimental results are reported and discussed.

4.3.1 Evaluation Metric

To evaluate the performance of the proposed APE approach, the two standard evaluation metrics: BLEU[7] which measures the degree of n-gram match between the model hypotheses and its target, and TER[8] which measures the number of edits required to change a system output into one of the references, are used in this work. Alignment Error Rate (AER) [Vilar et al., 2006] is used to evaluate the performance of alignment models.

4.3.2 Datasets

Monolingual data are used for training the denoising models and creating the bilingual dictionary. We collected eight million Myanmar sentences from various sources, including textbooks, Myanmar local news, Myanmar Wikipedia, ALT train data [Riza et al., 2016], and CC100-Burmese dataset [Conneau et al., 2019]. For the English monolingual data, we used ten million English sentences which consists of ALT training data and randomly extracted sentences from WMT monolingual News Crawl datasets[9]. To tokenize English texts, Moses tokenizer is used. We used the UCSYNLP word segmenter[10] to segment Myanmar sentences.

Parallel corpus is used to train the baseline NMT systems and a bilingual dictionary. It is also used to train/fine-tune the word alignment models. We manually collected around 224 thousand parallel sentences. Data statistics of parallel corpus are shown in Table 3.1 (in Chapter 3).

4.3.3 Model Configuration

The next subsections describe the details about the architecture and training procedure of baseline MT systems and our models in APE approach.

Baseline MT Systems:

The performance of the proposed APE approach is evaluated with three different test sets. These test sets are generated by a simple Transformer-based NMT, a fine-tuned mT5 and the Google Translate. We used PyTorch version of the OpenNMT project, an open-source (MIT) neural machine translation framework [Klein et al., 2018] to train the Transformer based NMT. All of the Transformer experiments are run on NVIDIA Tesla P100 GPU by following the settings listed in Table 3.2 (in Chapter 3). We were constrained by computational resources to *mt5-base*, which has 580M parameters in total. The pre-trained *mT5-base* model is initialized using Hugging Face's AutoModelForSeq2SeqLM[11]. The AdamW optimizer [Loshchilov et al., 2017]

[7] We used an implementation of BLEU from https://github.com/moses-smt/mosesdecoder in our experiments.
[8] http://www.cs.umd.edu/~snover/tercom/
[9] http://data.statmt.org/news-crawl/
[10] http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/wsandpos.html
[11]https://huggingface.co/docs/transformers/model_doc/auto#transformers.AutoModelForSeq2SeqLM

with a learning rate of 5e−4 and transformer's get_linear_schedule_with_warmup[12] scheduler are used for fine-tuning the mT5 model on 8 epochs with batch size of 16 and 1000 training iterations between checkpoints. Parallel datasets used in this experiment are tokenized into sub-word units by using SentencePiece[13] and used them to train/fine-tune and validate the baseline NMT and mT5 systems.

Denoising Models:

For denoising models, we used the architecture of 6-layer Transformer encoder/decoder [Vaswani et al., 2017]. Sockeye [Hieber et al., 2017; Kim et al., 2018] is used for training the Denoising autoencoder[14]. For the denoising rewriter, we use the same tool (i.e., OpenNMT tool) and settings for training as used in the baseline NMT. The target-side monolingual texts are used to train the denoising models. ALT dev dataset is used as the validation data.

Word Alignment Models:

The pre-trained LaBSE [Feng et al., 2020] model is used to produce the cross-lingual word embeddings for extracting the word alignments' information. The proposed word alignment model is compared with the following baselines:

1) fast_align [Dyer et al., 2013]: It is a simple and fast unsupervised word aligner with reparameterization of IBM Model 2.
2) GIZA++ [Och and Ney 2003; Gao and Vogel 2008]: It is an implementation of IBM models. We used five iterations each for Model 1, the HMM model, Model 3, and Model 4 to train GIZA++ by following the previous work of [Zenkel et al., 2020].
3) AWE-SoME [Dou et al., 2021]: It is a neural word aligner. It exploited the multilingual BERT to produce contextualized word embeddings from which the word alignments are extracted. This model can be applied with or without fine-tuning on parallel data.

4.3.4 Experimental Results

This section first discusses about the main results of the proposed APE system based with two different denoising strategies: DA and DW on the translation output of the three baseline MT systems: NMT, mT5, and Google Translate. Then, the proposed alignment model is evaluated by comparing its performance with state-of-the-art models. Moreover, a series of qualitative analysis and ablation studies are conducted on the baseline NMT output to further validate the reliability of the proposed denoising models and to better understand the importance of data preprocessing in the word alignment task.

---

[12] https://huggingface.co/docs/transformers/main_classes/optimizer_schedules#transformers.get_linear_schedule_with_warmup

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

[14] https://github.com/yunsukim86/sockeye-noise

4.3.4.1 Main Result

Table 4.1 reports the overall results of the proposed APE system. There are two denoising models to train the proposed APE system as described in the Subsection 4.2.3.1 and Subsection 4.2.3.2: *DA* and *DW*. The BLEU and TER metrics are used to evaluate the performance of the APE systems. The experimental results show that both versions of denoising models (i.e., *DA* and *DW*) used in the APE system improve the quality of the texts generated by the baseline NMT system. Firstly, the APE model trained with the *DW* model showed to give at least +4% BLEU score and -16% TER score, respectively. When we train the APE system with the *DA* model instead of the *DW*, an additional gain around +1% BLEU score and -2% TER score is obtained. Additionally, we conduct more experiments to further validate the effectiveness of the proposed systems on the state-of-the-art MT outputs. We use the APE system as a postprocessor to the texts translated by the state-of-the-art systems: mT5 and Google Translate. According to the reported results, in both English-to-Myanmar and Myanmar-to-English directions, APE system trained with the *DA* model can still be effective on improving the quality of the texts generated by mT5 and Google Translate. However, APE system trained with *DW* model failed to improve the quality mT5 and Google Translate translations in English-to-Myanmar direction.

Denoising models are developed using the same Transformer architecture but are trained on the different versions of noisy datasets. In this work, the insertion/deletion/reordering noise types showed a promising performance while mitigating these noises by using the denoising autoencoder model, *DA*.

Table 4.1: Performance of APE models.

	English→Myanmar		**Myanmar→English**	
	TER↓	BLEU↑	TER↓	BLEU↑
NMT	79.713	8.11	75.238	11.59
+APE (with DA)	**65.317**	**13.53**	**58.315**	**16.24**
+APE (with DW)	66.472	13.14	59.471	15.96
mT5	65.524	13.49	60.325	15.49
+APE (with DA)	**62.732**	**14.87**	**56.841**	**17.83**
+APE (with DW)	67.141	13.41	59.816	15.81
Google Translate	73.874	9.64	62.157	15.39
+APE (with DA)	**63.495**	**13.92**	**57.853**	**16.98**
+APE (with DW)	74.281	9.51	60.174	15.74

4.3.4.2 Word Alignment Result

Multilingual sentence embedding models are powerful tools for encoding similar texts from different languages into a shared embedding space. These embedding models have been applied in various downstream NLP tasks, such as clustering, text classification, and parallel sentences mining, and achieving the state-of-the-art performance in almost all tasks. The existing sentence embedding approaches like MUSE[15] or LASER[16], require parallel training data for mapping a sentence from one language directly into another language to encourage consistency between the sentence embeddings. Recently, the pre-trained LaBSE model that leverages recent advances on language model pre-training, results into a state-of-the-art model for encoding similar texts from different languages into a shared embedding space. LaBSE model is pre-trained on a BERT-like architecture and fine-tuned on a translation ranking task using the two objectives: masked language modeling (MLM) and translation language modeling (TLM). In this work, instead of sentence embedding purpose, we exploited the pre-trained LaBSE model to produce word embeddings. If the source and target words have similar meaning, LaBSE model enables to encode these words into a shared embedding space.

Given a source and target sentence pair, we first encoded all words in each sentence using LaBSE word embeddings. Then, based on the similarity of their word embeddings, all possible aligned word pairs are extracted by the proposed word alignment model. As described in Subsection 4.2.1.1, the threshold value for word similarity is set to 0.5. Every source-target word pair with the cosine similarity scores greater than the threshold is extracted. To evaluate the performance of the alignment, we use the Alignment Error Rate (AER) metric.

Table 4.2: Performance of word alignment models.

Model	**Setting**	**AER↓**
fast_align	bilingual	35.4
Giza++	bilingual	30.9
AWE-Some	w/o fine-tuning	28.2
	bilingual	25.8
Ours	w/o fine-tuning	26.1
	bilingual	**24.5**

Table 4.2 reports AERs of the proposed word alignment model and other popular word alignment models. English-Myanmar ALT test dataset was used for performance evaluation. According to the results, the proposed LaBSE-based word alignment model achieved state-of-

[15] https://github.com/facebookresearch/MUSE
[16] https://github.com/facebookresearch/LASER

the-art performance over the baseline models. In the table, the best score is in **bold**. Remarkably, the alignments which are directly extracted from LaBSE model(i.e., without fine-tuning the LaBSE model) already achieve better performance than the popular statistical word alignment models: fast_align and GIZA++. To more investigate the performance of our model in the bilingual setting, we fine-tuned the LaBSE model using the existing parallel corpus. In bilingual setting, the proposed alignment model achieves the best performance.

4.3.4.3 Qualitative Analysis

In the previous Subsection 4.3.4.1, the main results of this work showed that APE system trained using the denoising autoencoder (*DA*) model performs better than the system trained with the target-to-target rewriting based denoising model (*DW*). In addition, we conducted a qualitative analysis to further validate the trustworthiness of the DA-based APE system. We analyzed the post-edited texts of the two APE models: *DA* and *DW*, which were trained on different noisy datasets. Some examples of *DA*-based post-edited texts and *DW*-based post-edited texts are presented in Table 4.3 and Table 4.4, from Myanmar-to-English and English-to-Myanmar directions, respectively. In these tables, TER scores in each *mt* row are calculated regarding *tgt*; boldface words in each *mt* indicate words that need to be corrected to match the human-translated reference sentence, *tgt* of target-side. The findings show that the raw translation of English-Myanmar NMT, *mt*, needs too much number words to be corrected, whereas *mt* post-edited by the APE systems requires fewer corrections. Among the two APE systems, the text post-edited with *DA-based APE system* needs the fewest corrections. It means that DA-based APE system can correct the errors in *mt* and transform it to become a more accurate and fluent sentence, which is in turn similar to that of the reference sentence, *tgt*.

Table 4.3: Qualitative analysis for each Myanmar-to-English APE model trained in different denoising setting.

src	ဆစ်ဒနီ က ရန့်ဝစ်(ခ်) မြင်းပြိုင်ကွင်း မှ မျိုးသန့် ပြိုင်မြင်း ရှစ်ကောင် ဟာ မြင်းတုတ်ကွေးရောဂါ ကူးစက်ခံခဲ့ရတယ် ဆိုတာ အတည်ပြုခဲ့ပါတယ်။
tgt (=*ref*)	It has been confirmed that eight thoroughbred race horses at Randwick Racecourse in Sydney have been infected with equine influenza.
mt	Sydney **had** confirmed that **the** eight **more active consumer countries had** been infected with **the** influenza **virus**. (TER = 71.43)
mt (*post edited with DA*)	Sydney **had** been confirmed that **the** eight **genetic** race horses **had** been infected with **the** influenza at Randwick. (TER = 47.62)
mt (*post edited with DW*)	Sydney **had** been confirmed that **the** eight **genetic** race horses **had** been infected with **the** influenza. (TER = 52.38)
src	အန်အက်(စ်)ဒဗလျူ နှင့် ကွင်း(စ်)လန်း(ဒ်) တလျှောက်မှ အပန်းဖြေရာသုံးသော မြင်းများ ဒါဇင်များစွာ ကူးစက်ခံရ သော်လည်း ဒီဖြစ်ရပ် ဟာ ပြိုင်မြင်းများ အတွက် ပထမဆုံး ကူးစက်ခြင်း ဖြစ်ပါသည်။
tgt (= *ref*)	The cases are the first infections of race horses, despite infecting dozens of recreational horses across NSW and Queensland.
mt	The **incident is** the first **release** of dozens **of resorts throughout the trying season and is infected with for the first time**. (TER = 85.00)
mt (*post edited with DA*)	The **incident is** the first infections of dozens of **resorts** horses **throughout the** NSW and Queensland, race horses **are infected for the first**. (TER = 70.00)
mt (*post edited with DW*)	The **incident is** the first infections of dozens of **resorts** horses **throughout the** Queensland, race horses **are infected with for the first.** (TER = 75.00)

Table 4.4: Qualitative analysis for each English-to-Myanmar APE model trained in different denoising setting.

src	It has been confirmed that eight thoroughbred race horses at Randwick Racecourse in Sydney have been infected with equine influenza.
tgt (=*ref*)	ဆစ်ဒနီ က ရန့်ဝစ်(ခ်) မြင်းပြိုင်ကွင်း မှ မျိုးသန့် ပြိုင်မြင်း ရှစ်ကောင် ဟာ မြင်းတုတ်ကွေးရောဂါ ကူးစက်ခံခဲ့ရတယ် ဆိုတာ အတည်ပြု ခဲ့ပါတယ်။
mt	ဆစ်ဒနီ **တွင် ရောဂါ လက္ခဏာ ဖြင့် ဇူမား** မှ ပြိုင်မြင်း **ရှစ်စီး တုတ်ကွေးရောဂါ ကူးစက် ခံခဲ့ရသည် ဟု** အတည်ပြု **ခဲ့ကြသည်** ။ (TER = 80.00)
mt (*post edited with DA*)	ဆစ်ဒနီ **တွင်** ရန့်ဝစ်(ခ်) မြင်းပြိုင်ကွင်း မှ ပြိုင်မြင်း **ရှစ်စီး တုတ်ကွေးရောဂါ ကူးစက် ခံခဲ့ရသည် ဟု** အတည်ပြု **ခဲ့ကြသည်** ။ (TER = 53.33)
mt (*post edited with DW*)	ဆစ်ဒနီ **တွင်** မြင်းပြိုင်ကွင်း မှ ပြိုင်မြင်း **ရှစ်စီး တုတ်ကွေးရောဂါ ကူးစက် ခံခဲ့ရသည်** ။ (TER = 66.66)
src	The cases are the first infections of race horses, despite infecting dozens of recreational horses across NSW and Queensland.
tgt (= *ref*)	အန်အက်(စ်)ဒဗလျူ နှင့် ကွင်း(စ်)လန်း(ဒ်) တလျှောက်မှ အပန်းဖြေရာသုံးသော မြင်းများ ဒါဇင်များစွာ ကူးစက်ခံရ သော်လည်း ဒီဖြစ်ရပ် ဟာ ပြိုင်မြင်းများ အတွက် ပထမဆုံး ကူးစက်ခြင်း ဖြစ်ပါသည် ။
mt	**ထို ဖြစ်ရပ်များသည်** အန်အက်(စ်)ဒဗလျူ နှင့် ကွင်း(စ်)လန်း(ဒ်) တလျှောက်မှ **မြင်း** ဒါဇင်များစွာ အပန်းဖြေ **နိုင်** သော်လည်း **၊ ပြိုင်ပွဲ ၏** ပထမဆုံး **ရောဂါ ကူးစက်မှုများ ဖြစ်သည်** ။ (TER = 76.47)
mt (*post edited with DA*)	**ထို ဖြစ်ရပ်များသည်** အန်အက်(စ်)ဒဗလျူ နှင့် ကွင်း(စ်)လန်း(ဒ်) အပန်းဖြေ တလျှောက်မှ **မြင်း** ဒါဇင်များစွာ ကူးစက်ခံရ သော်လည်း **၊** ပထမဆုံး **ဖြစ်သည်** ။ (TER = 64.70)
mt (*post edited with DW*)	**ထို ဖြစ်ရပ်များသည်** အန်အက်(စ်)ဒဗလျူ နှင့် ကွင်း(စ်)လန်း(ဒ်) တလျှောက်မှ **မြင်း** ဒါဇင်များစွာ ပထမဆုံး **ကူးစက်မှုများ ဖြစ်သည်** ။ (TER = 70.58)

4.3.4.4 Ablation Study

Denoising Autoencoder:

As an ablation study, we tuned each parameter of the noise type and combined them incrementally to examine the effect of each noise type in the denoising autoencoder while post-editing the baseline English-to-Myanmar NMT translations. The results with different values of denoising parameters are presented in Table 4.5. Firstly, the reordering noise is tuned with

different values of d_{per}. A significant improvement in BLEU score was reached when the value of d_{per} is set to 5 since a local reordering usually involved a sequence of 5 to 6 words. When we tried to train with the value d_{per} greater than 5, we discovered that it cannot handle the long-range reordering because too many consecutive words are shuffled together, yielding no more improvement.

Next, we tuned the parameter of the deletion noise. When we used p_{del} = 0.1, it offered +1.16% BLEU score. However, the performance immediately degraded with a larger value of p_{del}. This is because it was difficult to see one-to-many in the similar target word substitution more than once in each sentence pair. Finally, we tuned the parameter of insertion noise. In this case, the best performance, +1.92% BLEU score was achieved with V_{ins} = 10. In general, increasing the V_{ins} value was not effective because it provided too many variants in the inserted word; it might not be related to its neighboring words.

Table 4.5: APE results with different values denoising parameters for Myanmar-to-English NMT

d_{per}	p_{del}	V_{ins}	**BLEU**
2			12.64
3			13.02
5			**13.16**
5	0.1		**14.32**
	0.3		13.85
5	0.1	10	**16.24**
		50	15.97
		500	15.32
		5000	14.86

Word Alignments:

In this work, we investigated the performance of the two pre-trained embedding models: mBERT and LaBSE, on the task of extracting word alignments. mBERT is a Transformer model pre-trained on the huge amount of multilingual Wikipedia texts with the object of masked language modeling (MLM). In our alignment extraction task, we used the 8-layer of mBERT word embeddings by following the work of Dou et al., [2021] and 12-layer of LaBSE embeddings, respectively. We further conducted the experiments with sub-word level and word level embeddings to investigate how the performance of word alignment varies with different levels of word embeddings.

The results on English-Myanmar language pairs are reported in Table 4.6. The LaBSE based word alignment model can notably outperforms the mBERT based model by a large margin in AER score. Although both mBERT and LaBSE models support both Myanmar and English languages in a single model, the word embedding vectors spaces of mBERT between

languages are not aligned, (i.e., the text with the same content in different languages would be mapped to different locations in mBERT word embedding the vector space). This finding suggests that LaBSE model is indeed helpful in the word alignments extraction task even in a low-resource setting where no training data is available. Moreover, we additionally examined the performance of both mBERT and LaBSE models in the sub-word level embeddings. For sub-word level embeddings, we segmented source and target texts into the sub-word units using the SentencePiece tool. According to the result, the performance of sub-word level embeddings is worse than the word level embeddings in both mBERT-based and LaBSE-based word alignment model. In sub-word level embeddings, for the short sub-word tokens, the context they potentially met during the embedding training was much more various than a complete word, and thus a direct translation of such token to a sub-word token of another language would be very ambiguous. We found that the word-to-word similarity calculation with cross-lingual embeddings depends highly on the frequent word mappings. Moreover, learning the mapping between rare words does not have a positive effect. Based on this finding, we leveraged LaBSE word embeddings without considering sub-word level in the proposed APE system.

Table 4.6: Alignment results with different word embeddings.

	Embedding Model	**Setting**	**AER↓**
word level	mBERT	w/o fine-tuning	35.7
		bilingual	26.4
	LaBSE	w/o fine-tuning	26.1
		bilingual	**24.5**
sub-word level	mBERT	w/o fine-tuning	38.6
		bilingual	36.8
	LaBSE	w/o fine-tuning	37.3
		bilingual	36.2

4.4 Related Work

Most recent automatic post editing (APE) studies primarily focus on the approaches to address and overcome APE training data sparsity problems. While recent research works have reported that augmenting the synthetic APE triplets ⟨*src*, *mt*, *pe*⟩ from parallel corpora using various noising schemes [Moon et al., 2021] and adding the augmented synthetic data to genuine data to expand the size of APE triplets [Dowmunt et al., 2016; Negri et al., 2018; Lee et al., 2020] can alleviate the APE data scarcity problem, other studies have highlighted several open challenges [Carmo et al., 2021]. The quality of the augmented synthetic data is a major challenge. These recent synthetic data augmentation works usually neglect to comply with minimum-editing criterion, where the post-edited data *pe* should be created by minimally editing *mt* yet maintaining the meaning of *src*. Therefore, the correction patterns detected in

their augmented data may differ from those occurring in the actual APE data, and thus it might limit the APE performance. Moreover, when the APE triplets are generated using the existing parallel data, training the baseline MT and APE model on the same data size is not effective [Chatterjee et al., 2018; 2019; Ive et al., 2020]. Besides, there is the fact that the post-edited sentence *pe* should not be a reference translation of MT corpus, since this would defeat the purpose of learning editing patterns for the MT output [Carmo et al., 2021]. This work considers the limitations in APE triplet augmentation and avoiding the absence of APE triplets that hinder the applicability of the APE task on English-Myanmar NMT. We pursue an alternative solution to develop an APE system without using any human-edited APE triplets.

Mainly, APE systems are used for improving MT output by utilizing information unavailable to the decoder and coping with systematic errors including adequacy and fluency errors of an MT system whose decoding process is not accessible. For this purpose, the work of Béchara et al., [2011] on English-French APE tried to keep the connection between MT output and its correspondence source sentence using word alignments in order to improve the adequacy. They first produced the new intermediate sentence by concatenating each word in MT output with "#" and aligned source word. Then, they trained the APE model to rewrite the intermediate sentence to the reference target sentence. However, their approach failed to improve on the English-to-French MT baseline and achieved only a small increase in BLEU of 0.65 absolute over its French-to-English baseline. The other work of Parton et al. [2012] also addressed the specific linguistic adequacy errors. They tried to correct the errors by either inserting or replacing words into the hypothesis. Their approach requires the three resources such as the phrase table, dictionaries and background MT corpus for only fixing certain word-choice errors (e.g. numbers, names and named entities). In our work, we consider on the same purpose but study an alternative approach that can be helpful even for very low-resource languages where APE triplets are unavailable. We propose a simple and effective APE pipeline with three main modules: (1)word alignments extraction module for detecting errors (unaligned target words) and missing information (unaligned source words) in the MT output with respect to its correspondence source sentence, (2) sentence enrichment module to enrich the MT output by removing errors and adding missing information, and (3) sentence denoising module to finally denoise the MT output to be in correct word and phrase order. This work demonstrates that an effective APE system can be developed without having access to the APE triplets.

A large and growing body of literature has investigated the use of pre-trained contextualized word embeddings derived from multilingually trained language models (LM) in the word alignments' information extraction task. Sabet et al. [2020] developed a neural word alignment model that aligns words using multilingual contextualized embeddings and achieved the good performance even in the absence of explicit training on parallel corpus. The recent work of Dou et al. [2021] also developed a neural word alignment model by leveraging the pre-trained mBERT embeddings and fine-tuned the model on the parallel corpus for achieving the better alignment results. The mBERT embeddings has impressive zero-shot cross-lingual transfer capabilities when fine-tuned on the various downstream NLP tasks. However, mBERT is not pre-trained with explicit cross-lingual supervision. Therefore, transferred performance can further be improved by aligning mBERT embeddings with cross-lingual signal. In our work, we use we use LaBSE embeddings instead of mBERT embeddings. LaBSE is a state-of-the-art sentence embedding model that encodes similar texts from different languages into a shared

embedding space and achieving competitive performance on the parallel sentence pairs extraction task. LaBSE model is helpful even on the low-resource languages for which there is no data available during training. The result of our experiments reported that LaBSE word embeddings is superior to mBERT in our proposed word alignments model.

The quality of the word-by-word translation text generated by a unsupervised MT system trained only on available monolingual corpora can further be improved with the denoising autoencoder [Kim et al., 2019]. Denoising autoencoder is a Transformer based model that takes a noisy sentence as input and produces a clean sentence as output, both of which are of the same language. Our proposed APE system used denoising autoencoder as the final postprocessor to denoise potential errors and word order in the input sentence.

In our APE approach, the target sentence enrichment module enriches the raw MT output by deleting systematic errors (i.e., deleting unaligned target words) and adding missing source-side information (i.e., finding the most similar target word for each unaligned source word and appending these target words to the end of the MT output). This process of unaligned source words to the closet target words substitution can be considered as the part of the word-by-word translation. By following the same idea as in word-by-word translation task, we also use the denoising model in our APE system to transform the enriched-version of MT output into a clean and fluent version. To investigate with different version of denoising models, we further design a denoising rewriter; a target-to-target rewriting model, as an alternative to the denoising autoencoder. We trained the denoising autoencoder (DA) and denoising rewriter (DW) with different settings of noises and then used each of them as a final module of the proposed APE system. According to the result of our experiments, the APE system trained with DA is effective for improving the quality of the texts generated not only by a simple Transformer-based NMT but also by state-of-the-art MT systems: mT5 and Google Translate.

4.5 Summary

To make the APE system more widely applicable for the most language pairs where APE triplets are unavailable, we propose a simple yet effective APE pipeline with three main modules. The proposed APE approach can correct the systematic errors in the texts generated by the current English-Myanmar NMT systems. According to the experimental results, our APE approach enables to significantly improve the quality of any machine-translated text in both English-to-Myanmar and Myanmar-to-English directions.

In this research, we identify three main principles underlying the recent successes in the absence of APE triplets. Then, we demonstrate how to apply these three principles to build an APE system without having the APE triplets. These three principles are word alignments information retrieval, sentence enrichment, and sentence denoising. Concisely, we first introduce a simple yet effective word alignment model that enables to extract alignments' information from LaBSE based contextualized cross-lingual word embeddings. The extracted word alignments' information is then used in sentence enrichment module where the semantic gap between the MT output and its corresponding source sentence is minimized by removing the detected errors (i.e., unaligned target words) and inserting missing source-side information (i.e., unaligned source words) in target sentence. Finally, the enriched version of the MT text is denoised to be more

fluent by the proposed denoising modules: denoising rewriter and denoising autoencoder. After passing denoising module, a clean and fluent sentence in the target language is obtained. According to the experimental results, the proposed APE system integrated with the above three principles give a promising performance even in the absence of APE triplets.

This work is the first attempt of APE research in a low-resource English-Myanmar language pair, where APE triplets are unavailable. Our findings in this work can encourage and help to do further APE research along this direction. The models used in this APE systems can be trained in both low and rich-resource settings, and can also be applied not only in the APE task but also in other MT-related works.

Chapter 5

Building Parallel Corpus from Monolingual Target Texts

5.1 Introduction

Building a high-quality machine translation (MT) system requires high-quality large-scale parallel corpus. A Parallel corpus consists of original texts in one language L_1 and their translations into other language L_2. The more of high-quality parallel data are available, the better the quality of NMT systems. However, many existing large-scale parallel corpora are limited to specific languages and domains. The vast majority of language pairs including English-Myanmar have very little, if any, parallel corpora due to the cost of their creation. In contrast, large amount of monolingual corpora are easier to obtain.

While a large body of literature has studied the use of additional parallel sentences created from comparable corpora for training machine translation (MT) systems is often an effective approach when dealing with low resource language pairs, where parallel corpora scarce. Findings in the literature show that there are two approaches that support the automatically creation of parallel sentences. The first approach is to extract parallel sentence pairs with a significant degree of parallelism form available topic-aligned parallel documents, called *comparable corpora* [Gre´goire and Langlais, 2018; Hangya and Fraser, 2019]. The second approach is to use a back-translation (BT) model trained on existing parallel data for creating new pseudo parallel corpus from available target monolingual sentences [Xu et al., 2019].

Myanmar is a low-resource language and thus only a small amount of English-Myanmar parallel data is currently available to train the MT models. The topic-level aligned documents (i.e. comparable corpora) that contains an amount of the mixture parallel and partially parallel English-Myanmar sentences are also not yet available. However, a plenty of monolingual English texts in various domains can be obtained easily. These monolingual English texts can be automatically backward translated into Myanmar texts to obtain additional parallel sentences for training the MT models.

This work investigates a simple yet effective approach that leverage target monolingual texts for improving the translation quality of both source-to-target MT model (a.k.a. a forward model) and target-to-source MT model (a.k.a. a backward model). We first construct a pseudo parallel dataset from English monolingual texts by utilizing the back-translation mechanism and further extract only the high-quality sentence pairs from the constructed dataset. The extracted parallel sentence pairs are used as the additional training data to train the new MT systems.

Back-translation is the state-of-the-art data augmentation approach to expand the size of training data. If we have a large amount of monolingual target texts, we can create a large-scale synthetic parallel data by using back-translation. However, there is no guarantee of the quality of the back-translated data. If the baseline MT model used in back-translation process is trained

on low-resource parallel data, it might have many noisy target translations in the back-translated texts. The quality of the training data plays an essential role in training standard statistical MT (SMT) and NMT systems. Mainly, NMT systems are very sensitive to noise in the inputs. Therefore, when we use the constructed pseudo parallel corpus to train NMT systems without filtering out the low-quality noisy parallel sentences, it may lead the NMT systems to the performance degradation. Regarding the noisy sentence pairs filtering approaches, a simple but effective approach called *Construct-Extract* that extracts only the high-quality parallel sentences from the constructed pseudo parallel corpus is proposed in this work. The proposed extraction approach is based on the sentence-level cosine similarity scores of any two sentence embedding vectors, i.e., the vector representations of the English target sentence and its correspondence back-translated Myanmar source sentence. A Siamese BERT-Networks and an additional MEAN pooling layer is used to calculate the sentence embedding vectors of each sentence pair.

The contribution of this work is as follows:

- We propose a neural-based parallel data creation framework called Contruct-Extract and demonstrate the feasibility of improving performance on the SMT and NMT tasks by the proposed framework.
- There are two main outcomes of the proposed framework. First, the pseudo parallel corpus that is constructed from English (target language) monolingual texts throughout the back-translation process in which an improved English-to-Myanmar backward model is applied. Second, the high-quality parallel sentences that are extracted from the constructed pseudo parallel corpus and can be used to supplement the training data (parallel corpus).
- The efficacy of the constructed-extracted parallel sentences on enhancing the performance of the final MT models is demonstrated with the experiments on English-Myanmar bidirectional translation tasks.

5.2 Related Work

To address data sparsity problem in machine translations, several research works leverage monolingual data to build the pseudo-parallel data for improving translation quality. Bertoldi and Federico [2009] addressed the domain adaptation problem by generating pseudo-parallel corpus from a monolingual in-domain corpus and use it as an additional training data. Niu et al. [2018] addressed low-resource and out-of-domain problems by applying the combination of back-translation and multilingual NMT system. Sennrich et al. [2016], Zhang et al. [2018], and Chen et al. [2019] obtained substantial improvements by using the pseudo-parallel corpus created from monolingual target and/or source texts via back-translation and/or self-training. as additional training data. Similarly, in this research work, we constructed a pseudo-parallel corpus from target monolingual texts and used it as the additional training data. Contrastingly, we applied automatic noisy sentence pairs filtering approach to obtain high-quality pseudo-

parallel corpus. The proposed system extracts only high-quality sentence pairs from the constructed pseudo-parallel corpus. We conducted MT experiments on low-resource English-Myanmar language pair to demonstrate the accuracy of the extracted pseudo-parallel corpus. Previous works [Xu and Koehn, 2017; Artetxe and Schwenk, 2019; Junczys-Dowmunt, 2018; Chaudhary et al., 2019] on parallel corpus filtering requires large clean parallel corpora or dictionaries and thus it performs poorly in our low-resource scenario.

The work of Imankulova et al. [2017] uses the Round Trip BLEU score between true and the synthetic sentences to filter noisy sentence pairs. I our work, we use sentence embeddings-based cosine similarity score instead of the Round Trip BLEU score.

The resent work [Jaiswal et al., 2020] applies a naive filtering model based on sentence similarity to filter out noisy pseudo parallel data. They apply MUSE (Multilingual Universal Sentence Encoder) [Yang et al., 2019] to obtain the sentence embeddings. In our work, we use Sentence-BERT (SBERT), a modification of the pretrained BERT network that use siamese and triplet network structures that derive semantically meaningful sentence embeddings and can be compared using cosine-similarity.

5.3 Construct-Extract: A Neural-based Framework for Building Bilingual Corpus

The proposed neural-based framework for creating a high-quality parallel corpus comprises the two major modules for (1) pseudo parallel corpus construction task and (2) high-quality parallel sentences pairs extraction task. Figure 5.1 shows an overview of the system containing these components. Briefly, the first module which generates a pseudo parallel corpus through a back-translation process is depicted by Figure 5.1 (a). The second module which applies the Siamese BERT-Network based architecture for extracting the high-quality from the corpus generated by the first module is depicted by Figure 5.1 (b).

For Figure 5.1 (a), we construct more parallel trans-lated texts through back-translation [Sennrich et al., 2016a] using a volunteer translator, i.e. an automatic back-translation of the 150k in-domain target mono-lingual English text into the source Myanmar language using pre-trained English-to-Myanmar back-ward MT model. To select a volunteer backward translator, we conduct experiments on the choice of SMT and NMT with the available parallel datasets. As a result, NMT generates more accurate and fluent translation outputs than SMT in both directions; thereby we choose it as our choice in the pipeline.

It might have the noisy target translations in Myanmar language in our constructed pseudo parallel corpus that is because the low-quality MT model trained on low-resource parallel data is used in the back-translation process. NMT system are very sensitive to noisy input data compared to SMT system. In some works, training the NMT systems using back-translated corpus as additional training data will cause the translation performance to deteriorate [Du and Way, 2017] or the translation performance will not be as good as expected. To eliminate this problem, we propose a simple and effective high-quality sentence pairs extraction model that incorporates a Siamese BERT networks based model with cosine similarity. The whole process of high-quality sentence pairs extraction task is shown in Figure 5.1 (b), in which Siamese BERT network is applied to indicate similar sentence embeddings between sentence vectors *u*,

v with cosine similarity to threshold only good quality sentence pairs. If the similarity score is greater than or equal to a decision threshold *p*, we add that pair into the training data as a good quality sentence pair.

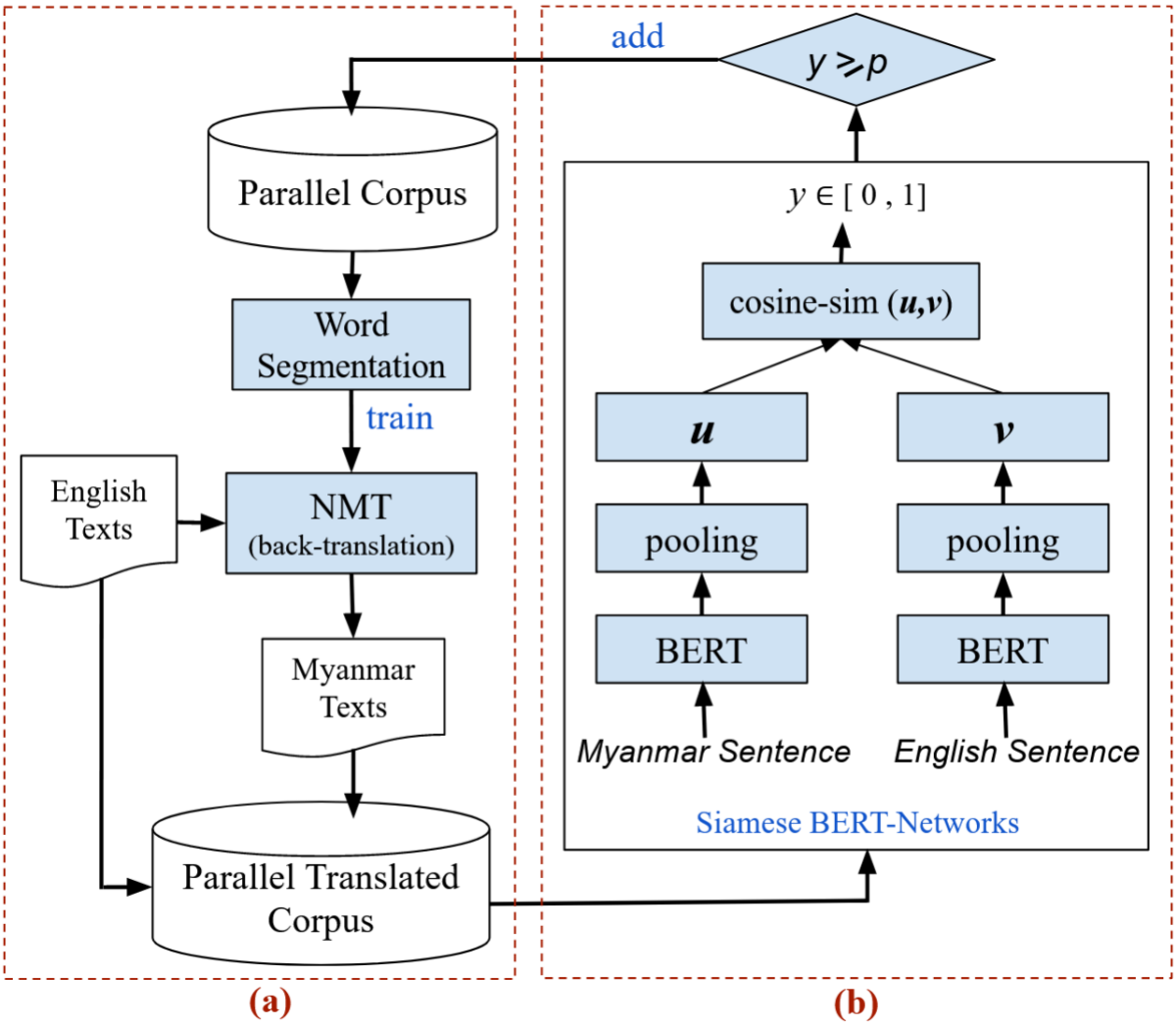

Figure 5.1: The proposed Construct-Extract framework for Myanmar-English parallel corpus creation.

5.3.1 Back-Translation

Back-translation [Sennrich et al., 2016a] is an effective approach for data augmentation by leveraging target monolingual text. Pseudo parallel corpus is generated throughout the back-translation process and that corpus can be used as additional data to train the MT systems. In this research, we tried the back-translation to construct parallel translated sentence pairs from collected monolingual English texts. In this work, a back-translation process is conducted to construct a pseudo parallel corpus (150k Myanmar-English parallel sentence pairs in total) by leveraging the collected 150k English (target language) monolingual sentences. Firstly, we pre-train an English-to-Myanmar NMT model on the existing parallel corpus and then use that

model in the back-translation process. However, the quality of the pre-trained model is not good enough to construct the high-quality pseudo parallel corpus because of the low-resource training data. It might contain noisy data (low-quality back-translated sentences) in the constructed corpus, which can severely impact the performance of NMT systems. Therefore, we design an extraction module where all of the noisy sentence pairs are removed from the constructed corpus.

5.3.2 Sentence Embeddings

Sentence-BERT (SBERT) [Reimers and Gurevych, 2019], a modification of the pretrained BERT network that uses Siamese and triplet network structures [Schroff et al., 2015], has set a new state-of-the-art performance on various sentence classification, clustering and sentence pair regression tasks such as semantic textual similarity. There are an increasing number of pretrained models that currently support more than 100 languages including English and Myanmar. Because these models are pre-trained based on the idea that a translated sentence should be mapped to the same location in the vector space as the original sentence, they can generate aligned vector spaces for every similar text in different languages (i.e., similar inputs in different languages are mapped closely in a vector space) [Reimers and Gurevych, 2020].

A pre-trained model, *distilbert-multilingual-nli-stsb-quora-ranking*, is used in our experiments to obtain the aligned vector spaces for every similar pair of English sentnce and its correspondence back-translated Myanmar sentence. Then, the cosine similarity is applied to indicate how much the input sentence pair is semantically similar to each other. In our experiments, we threshold the similarity between each sentence pair at 0.77. The model decides to add the pairs that have the similarity scores greater than or equal to the threshold into the existing training corpus as the good quality parallel sentence pairs. From Myanmar-English pseudo parallel corpus that contains 150 thousands sentence pairs in total, our model extracted only 92,111 sentence pairs based on the defined threshold value. We conduct the experiments to evaluate the effectiveness of our approach on the SMT and NMT tasks. We elaborate this in more details in the next section.

5.4 Experiments

This section describes the datasets and configuration of baseline MT systems that we have used in this work. Then, the results and qualitative analysis are discussed in detail.

Datasets:

For parallel data, there are around 224 thousand manually created and collected English-Myanmar parallel sentence pairs including parallel sentences from text books, Myanmar local news, and the ALT Corpus [Riza et al., 2016] for training the baseline MT systems. The development and test datasets we used in these experiments are only from the ALT corpus. Data statistics are shown on Table 3.1 (in Chapter 3). For creating a pseudo parallel corpus,

150k English sentences are collected from the internet. These monolingual English sentences are nearly in the same domain of the ALT corpus.

Model Configuration:

We evaluated the effectiveness of the proposed approach by performing machine translation experiments. We use phrase-based SMT (PBSMT) and Transformer-based model for the baseline SMT and NMT systems, respectively. We used Moses toolkit [Koehn et al., 2007] and GIZA++ [Och and Ney, 2003] to implement the word alignment process and to train PBSMT system. We applied grow-diag-final for phrases extraction and msd-bidirectional-fe heuristic for lexicalized word reordering. We used the default parameters of Moses to tune PBSMT model. The 5-gram language models with Kneser-Ney smoothing using KenLM [Heafield et al., 2013] are trained on Myanmar and English monolingual texts. The Transformer-based NMT models is trained using the PyTorch version of the OpenNMT toolkit, an open-source (MIT) neural machine translation framework [Klein et al., 2018]. The Transformer experiments were run on NVIDIA Tesla P100 GPU with the following parameters listed in Table 3.2 (in Chapter 3). To segment the Myanmar sentences, we used the Myanmar word segmenter proposed by Zin et al. [2021].

5.4.1 Main Result

The first module of the proposed framework constructs pseudo-parallel corpus (a.k.a constructed corpus) which contains 150 thousands parallel sentences via back-translation process. After adding these parallel sentences to the existing parallel corpus, we have more training data to train SMT and NMT systems. In general, the more training data helps the MT systems to obtain the better performance. However, NMT system has an issue with the low-quality noisy sentences. It might contain low-quality sentence pairs in the pseudo parallel corpus. For investigating and eliminating this problem, we further proposed a Siamese-BERT networks based extraction module that extract only 92,111 high-quality sentence pairs (a.k.a extracted corpus) from the 150k pseudo parallel corpus.

Table 5.1 and Table 5.2 illustrates the quality of the parallel corpus created by our Construct-Extract approach on the MT experiments. In these tables, the BLEU scores of SMT and NMT systems on three different data size settings: only on existing corpus, on existing corpus plus constructed data (all back-translated sentence pairs), and on existing corpus plus extracted data (high-quality sentence pairs from constructed data) are reported. In both directions, SMT systems gain an increase on the performance with more additional data. On the other hand, NMT systems trained using the extracted pseudo-parallel corpus as additional data returned the best translation performance. These findings suggest that translation accuracy of the NMT systems depends on both the size and quality of the training data. In this scenario, the proposed Construct-Extract mechanism can be the most useful for obtaining an improved pseudo-parallel corpus.

Table 5.1: BLEU scores for English-to-Myanmar MT systems.

Training Data	Total Sentences	SMT	NMT
Existing Parallel Corpus	204,535	7.63	8.11
+Constructed Corpus	+150,000	8.92	8.37
+Extracted Corpus ($p \geq 0.77$)	+92,111	8.61	8.51

Table 5.2: BLEU scores for Myanmar-to-English MT systems.

Training Data	Total Sentences	SMT	NMT
Existing Parallel Corpus	204,535	9.19	11.59
+Constructed Corpus	+150,000	9.43	12.21
+Extracted Corpus ($p \geq 0.77$)	+92,111	9.38	12.41

5.4.2 Qualitative Analysis

This section describes the qualitative analysis of the sentence pairs constructed by our approach. The accuracy of the three English-Myanmar sentence pairs are presented in Figure 5.2. The cosine similarity score of each pair is also provided. To facilitate a comparison with the original English sentences, the back-translated Myanmar sentences have been translated into English language using Google Translate tool. Among these three example pairs, the first two pairs have similarity score more than 0.77 (the pre-defined threshold value) are extracted as high-quality pairs. We conducted many experiments with different threshold values to obtain the best one for deciding whether the sentence pair is similar or not. BLEU scores of MT systems on different threshold values are presented in Figure 5.3. In both English-to-Myanmar and Myanmar-to-English directions, the best BLEU scores are obtained while we used 0.77 as a threshold value.

Sentences from English Monolingual Data	Model generated Parallel Myanmar Sentences	Translation of Model generated Myanmar Sentence into English for Comparison (Google Translate)	Model generated Similarity Score
The Metro is currently used by 700 million passengers per year .	မက်ထရို ကို တစ်နှစ် လျှင် ခရီးသည် ၇၀၀ သန်း လောက် အသုံးပြု နေသည် ။	Metro is used by about 700 million passengers a year.	0.8203
It will help tourism and other economic development in the Czech Republic .	၄င်းသည် ချက် သမ္မတနိုင်ငံ တွင် ခရီးသွား နှင့် အခြား စီးပွားရေး ဖွံ့ဖြိုးမှု ကို ကူညီ လိမ့်မည် ။	It will help tourism and other economic development in the Czech Republic.	0.8127
A rough estimate is that about 300 SMEs will be able to benefit from investments in equity capital each year .	ညံ့ဖျင်း သော ခန့် မှန်းချက် တစ်ခု သည် နှစ်စဉ် အရင်းအနှီး များ တွင် ရင်းနှီးမြှုပ်နှံမှု မှ အကျိုးကျေးဇူး ရရှိ နိုင် လိမ့်မည် ဟု သတ်မှတ် ကြေး ကောက် ယူခြင်း သည် ။	A bad estimate is that an annual fee will be levied on the return on investment in capital.	0.7414

Figure 5.2: A sample of constructed sentence pairs (monolingual English sentences and their corresponding back-translated Myanmar sentences).

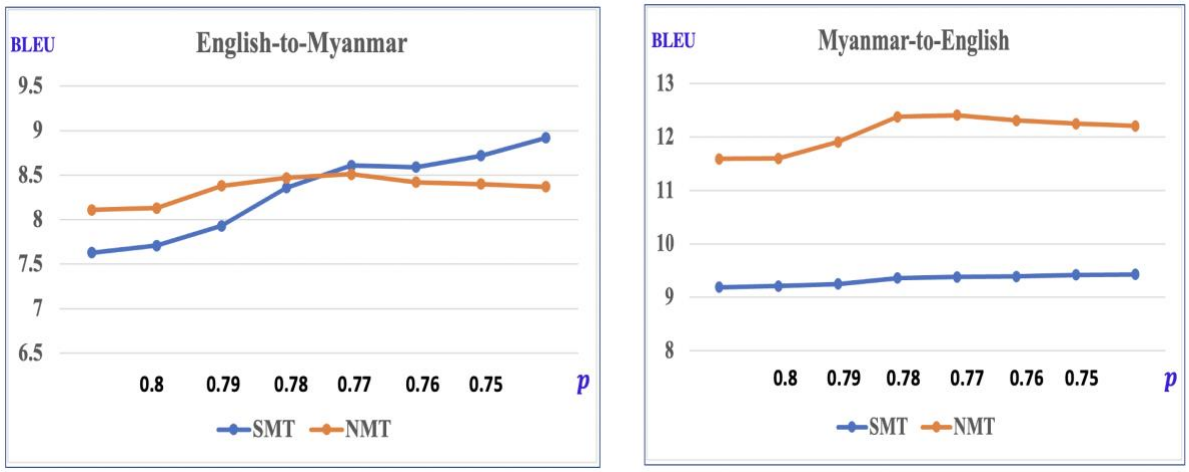

Figure 5.3: BLEU scores of English-Myanmar MT systems on different threshold values

5.5 Summary

This work is motivated by the expectation of improving the quality of translation systems in a low-resource setting. To achieve this goal, we propose a simple and effective data

augmentation and extraction approach as our main contribution. We used back-translation to augment pseudo parallel corpus. A Siamese-BERT-Networks-based architecture with addition mean pooling layer is used to generate sentence embeddings in our high-quality sentence airs extraction approach. The cosine similarity is applied to calculate the similarity score of the sentence pair. We performed SMT and NMT experiments to validate the performance of our proposed framework.

According to the experimental results, the constructed and extracted parallel datasets facilitate a significant improvement in the translation quality when compared to a generic system. In this English-Myanmar low-resource setting, our proposed framework is indeed beneficial for all SMT and NMT systems to obtain a remarkable percentage which increases in the BLEU scores. However, yielding a significant BLEU score, the constructed-extracted data is less effective to support the MT systems due to the lack of coverage on the sentence categories in the training and test datasets. The existing training and test data used in this work contains sentences from 13 different categories: crime and law, culture and entertainment, disasters and accidents, economy and busi- ness, education, the environment, health, obituaries, politics and conflicts, science and technology, sports, Wackynews, and weather. However, the collected monolingual English texts used in this work only cover 40 percent out of these categories. In the future, we plan to extend the proposed framework with a generative adversarial network for synthesizing high quality sentence candidates. We also plan to collect more monolingual texts in different categories to cover the categories of test dataset.

Chapter 6

Building Parallel Corpus from Comparable Corpora

6.1 Introduction

The use of parallel corpus extracted from the comparable corpora as the additional data for training machine translation (MT) systems is an effective approach when dealing with low resource language pairs, where parallel corpora are scarce [Smith et al., 2010]. Findings in the literature show that there are two approaches that support the automatically creation of parallel corpus. The first approach is creating parallel corpus by extracting the parallel sentences with a significant degree of parallelism from the available comparable corpora [Smith et al., 2010; Karimi et al., 2017; Azpeitia et al., 2018; Gregoire et al., 2018; Hangya and Fraser 2019; Steingrimsson et al., 2021; Althobaiti 2021]. The second approach is to use a back-translation (BT) or/and self-training (ST) for creating new pseudo parallel corpus from the available target monolingual data [Xu et al., 2019; Chen et al., 2019; Zin et al., 2021].

Increasing the size of training data is essential for improving low-resource machine translation (MT) systems. This work investigates an effective framework to create an additional parallel corpus from the available comparable corpora. The proposed framework consists of two main parts. The first part is to *expand* the size of comparable corpora and the second part is to *extract* parallel sentences from the expanded comparable corpora. While deal with the low-resource language pairs, the collected comparable corpora may contain only a small number of parallel sentences. Therefore, expanding the size of comparable corpora will be beneficial for obtaining many parallel sentences. We evaluate the impact of our approach on the English-Myanmar low-resource language pair.

The main contributions of this work are as follows:

- We first propose a bidirectional data augmentation approach by effectively utilizing self-training, back-translation, and the recent DbAPE [Zin et al., 2022] system, for expanding the data size of the collected comparable corpora. For every sentence in one language, we augment three aligned sentences in other languages using three different MT models. Having many aligned target (or source) sentences for a given source (or target, respectively) sentence, we can choose the best quality one.
- We improve a scoring formula for computing the semantic similarity of the two sentences. Our scoring formula can give the highest scores to the correctly aligned sentence pairs.
- We then propose an effective approach to extract parallel sentences from the expanded comparable corpora by applying our scoring formula on the LaBSE [Feng et al., 2020] based sentence embeddings.

- We quantitatively and qualitatively examine the main advantages and shortcomings of the parallel corpus created from comparable corpora by our approach when they are used as additional training data for MT.

The performance of NMT systems has been proven to mainly depend on not only the quantity but also the quality of the training data [Khayrallah and Koehn 2018]. Determining the quality and/or reliability of all of our created parallel corpus is a fraught and complex task, in part due to the lack of any single, widely accepted definition of what "parallel corpus quality" is. In our experiments, we create three types of parallel corpus and thus we have altogether too many numbers of sentence pairs (over 276k sentence pairs as shown in Table 6.6) for assessing the quality. Assessing the quality and/or reliability of these sentence pairs by bilingual subject-matter experts is relatively expensive and time-consuming, and thus we use BLEU as a proxy for human evaluation of quality. BLEU is a standard metric for evaluating the output quality of NMT systems. The higher the BLEU scores, the better the translations. Therefore, reliability and/or quality of the parallel corpus can be determined based on the improved BLEU scores of NMT systems trained on it. In this work, our experiments confirm that BLEU scores of NMT systems increased by about 6 points in both translation directions when our created parallel corpus is used as the additional training data.

The remainder of this chapter is organized as follows: Section 6.2 discusses previous studies related to this work. The detail about the data and our proposed system are explained in Sections 6.3 and 6.4. Experiments and results are discussed in Section 6.5. Finally, Section 6.6 presents the conclusion and the future work.

6.2 Related Work

To address training data sparsity in machine translations, comparable corpora have been shown to be a useful source for creating parallel data that can be used to supplement parallel corpus and can help to improve MT quality [Wolk et al., 2016; Hangya and Fraser 2019]. Afli et al. [2016] extracted English-French parallel sentences from a multimodal comparable corpus built from the Euronews[17] and TED[18] websites. Smith et al. [2010] and Chu et al. [2015] also extracted parallel data from Wikipedia and Ling et al. [2014] employed a crowdsourcing paradigm to obtain high-quality parallel data from Twitter.

Multilingual sentence embedding models achieve strong results in multilingual sentences similarity search even for low-resource language pairs. These models have also been applied to extract parallel sentences from comparable corpora, obtaining the state-of-the-art performance [Schwenk 2018; Artetxe and Schwenk 2019]. The recent work of Ramesh et al. [Ramesh et al., 2022] extracted parallel sentences from the web by using LaBSE model to produce multilingual sentence representations for aligning sentences and approximate nearest neighbor search for searching in a large collection of sentences. Steingrímsson et al. [2021] extracted parallel sentence candidates using an inverted index-based cross-lingual information

[17] https://www.euronews.com/
[18] https://www.ted.com/

retrieval (CLIR) tool called *FaDA* [Lohar et al., 2016], that requires a bilingual lexicon. The extracted sentence candidates were then scored using LaBSE embeddings-based similarity score and word alignment score. Recently, Althobaiti [2021] created parallel sentences from comparable corpora using a bilingual dictionary to translate English sentences into Arabic, and a word vectorization method such as TF-IDF to yield better results when computing similarity between sentences with mistakes in structures and syntax. Our work differs in that we first *expand* the low-resourced comparable corpora by creating many similar candidates of sentences in both directions using three pre-trained MT systems and DbAPE [Zin et al., 2022] system via self-training and back-translation processes (see Subsection 6.4.1.1 and Subsection 6.4.1.2 for a more detailed description of self-training and back-translation processes, respectively). And then, we *extract* parallel sentences from the expanded comparable corpora; we define a new scoring formula by exploiting the cosine similarity of the source-target sentence pair and the margin between the cosine and the average cosine of its k nearest neighbors in both directions. Hence, our approach does not require the bilingual lexicon for creating parallel sentences.

This work is an extension of the previous works of Zin et al. [2021; 2022]. Zin et al. [2021] described a method to effectively increase the size of training data for MT systems via filtering pseudo-parallel corpus obtained by using back-translation approach. Previously, Zin et al. [2021] leveraged only English monolingual texts for creating additional parallel corpus. To fill this gap, this work studies how comparable corpora are leveraged as an alternative approach for creating parallel corpus that can be supplemented to training data. Denoising-based automatic post-editing (DbAPE) system developed by Zin et al. [2022] is applied in this work to improve the augmented sentences' quality.

6.3 Data

In this section, we describe the detail of our training data and the applied preprocessing techniques.

6.3.1 Parallel Corpus

The parallel corpus consists of two datasets. For the first dataset, we manually collect and create 204,535 English-Myanmar bilingual sentences from various domains, including news articles and textbooks. The second dataset is the Asian Language Treebank (ALT) corpus [Thu et al., 2016] which consists of 18,082 training sentence pairs, 1,000 validation sentence pairs and 1,017 test pairs from English originating news articles. Data statistics are shown on Table 3.1 (in Chapter 3).

6.3.2 Comparable Corpora

We manually collect 150,000 English monolingual texts and 137,680 Myanmar monolingual texts from various domains: crime and law, disasters and accidents, economy and business, politics and conflicts, sports, international news, weather, and daily conversations. These

English and Myanmar monolingual texts contain a mixture of parallel and partially parallel sentences, and thus are used as the comparable corpora in this work. We denote English (***source***) monolingual dataset and Myanmar (***target***) monolingual dataset by M_S and M_T , respectively.

6.3.3 Data Preprocessing

The collected Myanmar monolingual data contains texts in both Unicode and Zawgyi encodings. The myanmar-tools[19] library is used to classify and convert all Zawgyi texts to Unicode. We tokenize English sentences using Moses [Koehn et al., 2007]. For tokenizing Myanmar texts, we used the Myanmar word segmenter developed by Zin et al. [2021].

6.4 Expand-Extract: A Parallel Corpus Mining Framework from Comparable Corpora

This section describes our proposed *Expand-Extract* framework for creating parallel corpus from comparable corpora under low-resource conditions. The *Expand-Extract* framework comprises two main modules for (1) *expanding* the data size of comparable corpora and (2) *extracting* parallel sentences from the expanded comparable corpora. Figure 6.1 shows an overview of the framework containing these components.

Briefly, Figure 6.1 (a) depicts the first *Expand* module which augments data bidirectionally (i.e., source-to-target and target-to-source directions) to expand the data size of the original comparable corpora with the synthetic data composed by the source sentences from source monolingual dataset with their corresponding three machine-translated target sentences and vice versa. Figure 6.1 (b) depicts the second *Extract* module which finds and extracts parallel sentence pairs from the expanded comparable corpora. Each of the resulting three datasets (***s2t𝒫***, ***t2s𝒫***, and ***bd𝒫***) is then employed as an additional training corpus.

6.4.1 Expanding the Data Size of Comparable Corpora

English-Myanmar is a low-resource language pair and thus the collected comparable corpora may contain only a small amount of parallel sentence pairs. Expanding the data size of the comparable corpora may be beneficial for obtaining a large amount of parallel data. Previous works [Ueffing 2006; Sennrich et al., 2015; Zhang and Zong 2016; He et al., 2019] applied the data augmentation approaches such as self-training, i.e., the forward (source-to-target) translation, and back-translation, i.e., the target-to-source, approaches to expand the size of parallel corpus by exploiting source and target monolingual texts. This work also applies these approaches to expand the data size of the low-resource comparable corpora for English-Myanmar translation.

[19] https://github.com/google/myanmar-tools

In the reminder of this work, we denote the forward (English-to-Myanmar) and backward (Myanmar-to-English) translation models by $\overrightarrow{f}$ and $\overrightarrow{g}$, respectively.

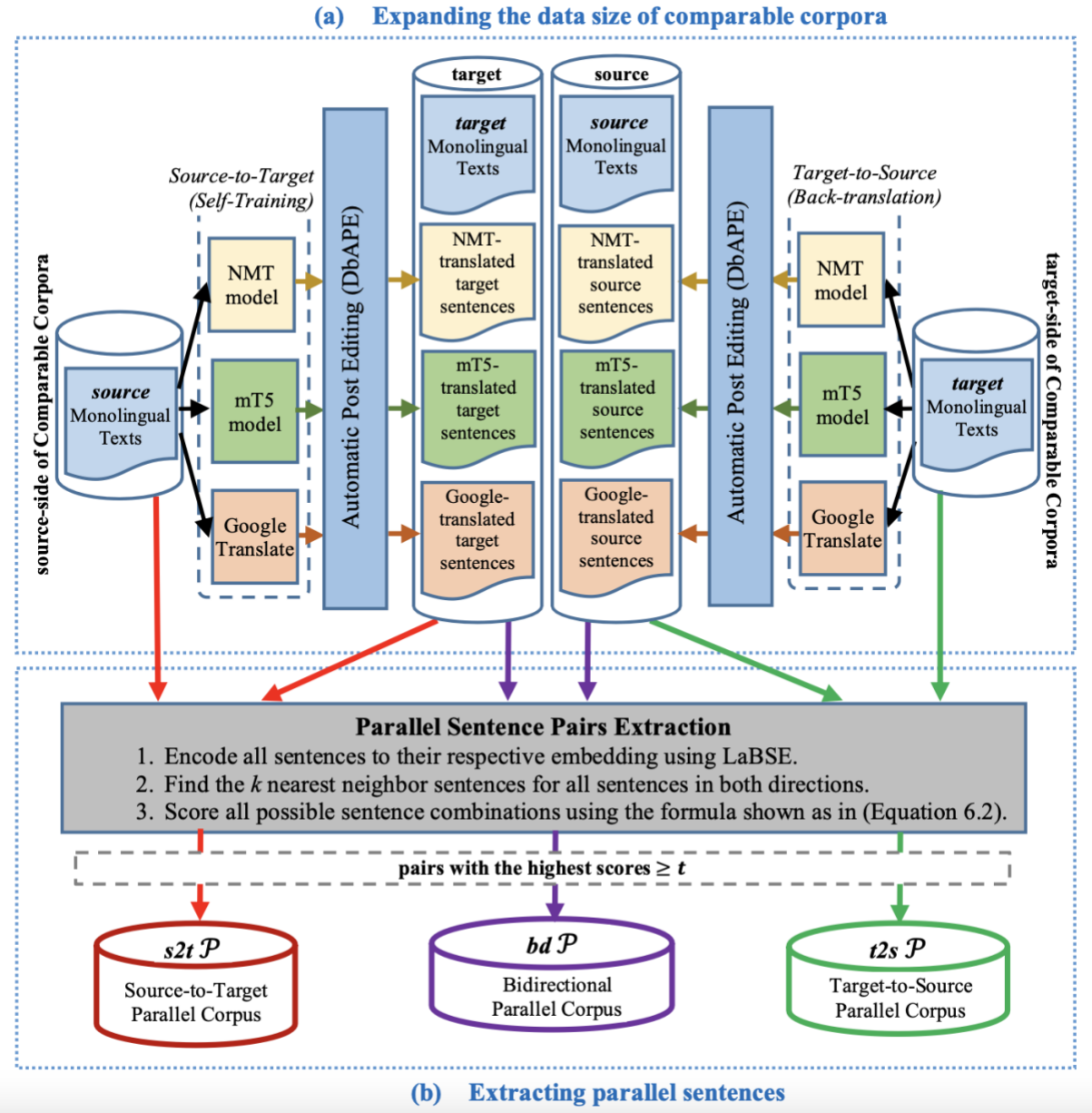

Figure 6.1: The proposed Expand-Extract framework for creating parallel corpus from comparable corpora.

6.4.1.1 Self-Training (ST)

Self-training [Ueffing 2006; Zhang and Zong 2016; He et al., 2019] is an effective data augmentation method leveraging source side monolingual texts. To perform self-training, we first train $\overrightarrow{f}$ on the existing parallel corpus (shown in Table 3.1 (in Chapter 3)) and use it to

translate $\mathcal{M}_S$ (150,000 English sentences) to produce synthetic target side data, denoted by $\overrightarrow{f}(\mathcal{M}_S)$. In this work, we use three forward translation models: NMT, mT5 and Google Translate. Therefore, after self-training (forward translation) process, we obtain three kinds of synthetic target side datasets (450,000 Myanmar sentences in total). These datasets are added to the target side of the original comparable corpora.

6.4.1.2 Back-translation (BT)

Back-translation [Sennrich et al., 2015] is a state-of-the-art data augmentation method leveraging target side monolingual texts. To perform back-translation, we first train $\overrightarrow{g}$ on on the existing parallel corpus (shown in Table 3.1 (in Chapter 3)) and use it to translate $\mathcal{M}_T$ (137,680 Myanmar sentences) to produce synthetic source side texts, denoted by $\overrightarrow{g}(\mathcal{M}_T)$. We train and use three back-translation models based on NMT, mT5 and Google Translate, and obtain three kinds of synthetic source side datasets (413,040 English sentences in total). These datasets are added to the source side of the original comparable corpora.

6.4.1.3 Automatic Post-editing (APE)

The APE system aims to correct the systematic errors in both forward and back-translated texts. For English-Myanmar low-resource language pair, in which only low-accurate machine translation systems can be used in self-training and back-translation processes, there is no guarantee of the quality of machine-translated synthetic texts. In order to improve the quality of the synthetic texts, we apply the denoising-based automatic post-editing (DbAPE) [Zin et al., 2022] system which automatically detects and corrects errors in the texts. DbAPE is a pipeline system consisting of three main modules. The first module is word alignment information retrieva1 that detects the systematic errors in a machine-translated text. Target sentence enrichment and target sentence denoising are the second and third modules of the DbAPE system. These modules transform any machine-translated sentence into an accurate and fluent one. DbAPE does not require APE data for training and thus can be used easily to train with any available monolingual data. The quality of the source and target sides synthetic datasets from comparable corpora have been shown to improve when applying the DbAPE.

6.4.2 Extracting Parallel Sentences

Extracting parallel sentences from comparable corpora is an effective approach to increase the training data for low-resource MT systems. Given two comparable corpora in different languages, the task is to find and extract the sentence pairs that are translations of each other. The multilingual encoder, LaBSE [Feng et al., 2020], has been reported to be competitive in extracting parallel sentences from comparable corpora by taking the nearest neighbor of each source sentence in the target side according to cosine similarity of their respective embeddings, and filtering those below a fixed threshold. However, the scale of cosine similarity is not being globally consistent across different sentences [Artetxe and Schwenk 2018]. For instance, Table 6.1 shows an example of where an incorrectly aligned sentence pair (i.e., **m2** and its nearest

neighbor) has a larger cosine similarity (i.e., 0.7977) than a correctly aligned one (i.e., **m1** and its nearest neighbor with cosine similarity score 0.7779). This is caused by the cosine similarity of different sentences being in different scales, making it a poor indicator of the confidence of the prediction. Therefore, using only cosine similarity to extract the sentence pairs through a fixed threshold is impossible to obtain the correctly aligned pairs. To tackles this issue, Artetxe and Schwenk [2018] considered the margin between a given candidate and the rest of the k nearest neighbors. They showed that the margin between the similarity of a given candidate and that of its k nearest neighbors is a better indicator of the strength of the sentence pair alignment. In their work, the margin between the cosine of a given candidate and the average cosine of its k nearest neighbors in both directions are considered as follows:

$$\text{margin\_score}(x, y) = \text{margin}(\cos(x, y), \sum_{z \in NN_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in NN_k(y)} \frac{\cos(y,z)}{2k}) \qquad (6.1)$$

where x and y are the source and target sentences, and $NN_k(x)$ denotes the k nearest neighbors of x in the other language, and analogously for $NN_k(y)$. We explore different margin functions: *absolute* ($\text{margin}(a, b) = a$), *distance* ($\text{margin}(a, b) = a - b$), and *ratio* ($\text{margin}(a, b) = a/b$), proposed by Artetxe and Schwenk [2018]. The *ratio* function yields the best results.

Table 6.1: Example of the nearest neighbors of the two Myanmar sentences (**m1** and **m2**) on comparable corpora along with their cosine similarities

(m1)	ကျွန်မ အတွက် ပို့ လိုက် တဲ့ ပန်းတွေ အတွက် ကျေးဇူးတင် ကြောင်း ပြော ချင် လို့ ဖုန်းဆက် လိုက် တာပါ ။ (Reference: I am calling to tell you how much I appreciate for the flowers you sent me.)
0.7779	I was just about to call to tell you how I appreciate the flowers you sent me.
0.5997	Thank you for your call.
0.5416	Thank you very much for inviting me.
0.5233	Thank you for calling Air France.

(m2)	လယ်ယာထွက်ကုန်များတွင် လက်ဖက်၊ ဆန်၊ သကြား၊ ဆေးရွက်ကြီး၊ ပရုတ်၊ သစ်သီး နှင့် ပိုးချည် တို့ ပါဝင်သည်။ (Reference: Agricultural products consist of tea leaf, rice, sugar, tobacco, camphor, fruits, and silk.)
0.7977	Main crops include wheat, sugar beets, potatoes, cotton, tobacco, vegetables, and fruit.
0.6407	The fertile soil supports wheat, corn, barley, tobacco, sugar beet, and soybeans.
0.6280	The important crops grown are cotton, jowar, groundnut, rice, sunflower, and cereals.
0.6247	Main agricultural products include grains, cotton, oil, pigs, poultry, fruits, vegetables, and edible fungus.

In this work, we conduct the experiments using not only cosine similarity but also margin-score. Although, margin-score function gives the better results than cosine similarity, considering both of them becomes the best function to indicate the correctly aligned sentence pairs. Therefore, our final scoring function is considered based on cosine similarity and margin-score as follows:

$$\text{score}(x, y) = \cos(x, y) + \text{margin\_score}(x, y) \quad (6.2)$$

When extracting parallel sentences, we perform the following steps to generate the candidates:

1. Each source sentence is aligned with exactly one best scoring target sentence.[20] Some target sentences may be aligned with none (this is because the number of source sentences is not equal to the number of target sentences in comparable corpora) or with multiple source sentences.
2. Each target sentence is aligned with exactly one best scoring source sentence. As Step 1, some source sentence may be aligned with multiple target sentences or with none.
3. The candidates generated from the previous steps are combined and the duplicated ones are removed.

These candidates are then sorted according to their scores (as in Equation 6.2), and a fixed threshold is applied to extract only the highly similar candidates.

6.5 Experiments

This section describes the configuration of the MT systems that we have used in this work. Then, the results, ablation study and analysis are discussed in detail.

6.5.1 Implementation Details

In the data augmentation process of expanding the data size of comparable corpora, we apply back-translation and self-training approaches using the pre-trained MT systems: Transformer [Vaswani et al., 2017] based NMT, mT5, and Google Translate. The parallel data shown in Table 3.1 (in Chapter 3) is used for pre-training the NMT system and fine-tuning the mT5 system. We utilized PyTorch version of OpenNMT toolkit developed by Klein et al. [2018] to train NMT system. The encoder and decoder consist of 6 layers, 8 attention heads, and the hidden size is kept to 512. We used the Adam optimizer with dropout, and the number of training step is 200,000. For mT5 system, we initialize the pre-trained mT5-base model using Hugging Face's AutoModelForSeq2SEqLM[21].

6.5.2 Expansion of Comparable Corpora

After applying the self-training, back-translation, and post-editing on the translated texts with the DbAPE system, the Expand module of our proposed framework bidirectionally expands the data size of existing comparable corpora with additional 413,040 English sentences and 450,000 Myanmar sentences as shown in Table 6.2. After mixing these augmented data with

[20] For efficiency, only the k nearest neighbors over cosine similarity are considered, where the neighborhood size k is the same as that used for the margin-based scoring shown as in Equation 6.1.
[21] https://huggingface.co/docs/transformers/model_doc/auto#transformers.AutoModelForSeq2SeqLM

the existing comparable corpora, we have more potential parallel sentences.

Table 6.2: Statistics of expanded comparable corpora

Language	Data Creation Approach	No. of Sentences	Total Sentences
English	Original (manually collected)	150,000	563,040
	Self-Translation (NMT, mT5, Google Translate) + post-edited by DbAPE	413,040	
Myanmar	Original (manually collected)	137,680	587,680
	Back-translation (NMT, mT5, GoogleTranslate) + post-edited by DbAPE	450,000	

6.5.3 Extraction of Parallel Sentences

To find and extract parallel sentences from the expanded comparable corpora, we first encode all sentences to their respective embeddings using the LaBSE model. Once we have the sentence embeddings, we find the k nearest neighbor sentences for all sentences in both directions. Then, we score all possible sentence combinations using Equation 6.2 (mentioned in Subsection 6.4.2). The pairs with the highest scores are most likely parallel pairs. For a good quality, a threshold t is used to extract the pairs above that threshold.

For English-Myanmar language pair, there is no test set previously made available for the parallel sentence extraction task. Therefore, for evaluation of our extraction approach, we created a training dataset with 50,000 English sentences and 48,500 Myanmar sentences in which 2,017 sentence pairs are aligned. The precision, recall and F1 scores on this dataset with k = 17 is shown in Table 6.3. The optimal threshold that leads to the highest F1-score is 2.16 when the typical choice for k is 17. To obtain the optimal value for k , we conduct experiments with different values. The F1 scores on different values of k are provided in Figure 6. 2, where we can see that a high accuracy is obtained by considering 17 nearest neighbors for each sentence in both directions.

Table 6.3: The result (Precision, Recall, and F1) on the English-Myanmar dataset used to optimize the threshold value for the parallel sentence extraction task

Language	Training Sentences	Gold Sentences	P	R	F1	Optimal threshold (t)
English	50,000	2,017	95.0	94.6	94.8	2.16
Myanmar	48,500	2,017				

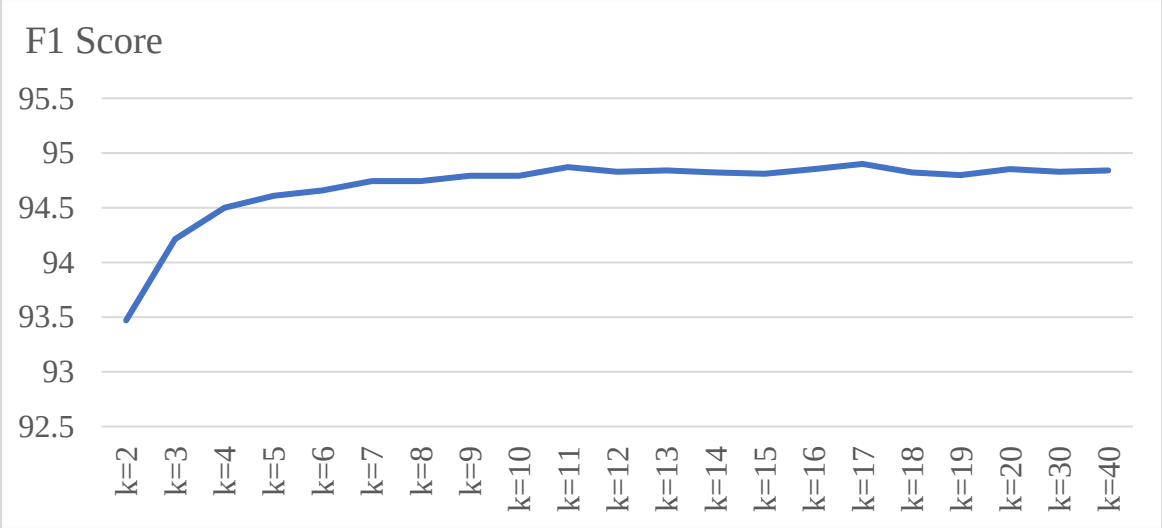

Figure 6.2: F1 scores on different values of k for extracting English-Myanmar parallel sentences.

To evaluate our scoring function, we compared ours with the previous state-of-the art methods. Table 6.4 reports precision, recall and F1 scores on three different scoring mechanisms: cosine similarity [Schwenk 2018], margin-score [Artetxe and Schwenk 2018], and ours (Equation 6.2). According to the results in Table 6.4, it is clear that our scoring function outperforms the previous mechanisms, and thus it can effectively apply in the parallel sentence pairs' extraction task. Indeed, we explore different values of k and scoring functions. The experimental results on English-Myanmar dataset show that using the value of $k = 17$ and the combination of cosine similarity and margin-based scoring yield the best results.

Table 6.4: The result (Precision, Recall, and F1) on different scoring mechanisms

Scoring Method	Precision	Recall	F1
Cosine similarity [Schwenk 2018]	89.3	87.5	88.4
Margin-score [Artetxe and Schewenk 2018]	93.9	94.5	94.2
Ours	**95.0**	**94.6**	**94.8**

To further evaluate our approach on different language pair, we used the training set of English-German language pairs from the BUCC 2018 shared task[22] . A comparison of our approach with the previous works of Feng et al. [Feng et al., 2020] and Artetxe and Schwenk [2018] is shown in Table 6.5. In the table, the precision, recall and F1 scores of each work are presented. According to the result, our framework with our scoring function (Equation 6.2)

[22] https://comparable.limsi.fr/bucc2018/bucc2018-task.html

outperforms the previous works.

Table 6.5: The result (Precision, Recall, and F1) on the BUCC English-German Training Set

Language	Training Sentences	Gold Sentences	Framework	P	R	F1
English	399,337	9,580	Artetxe and Schwenk [2018] (Cosine Similarity)	84.9	80.8	82.8
German	413,869	9,580	Artetxe and Schwenk [2018] (Margin-score)	**95.3**	94.4	94.8
			Feng et al. [2020] (Cosine Similarity)	92.3	92.7	92.5
			Ours	94.9	**94.9**	**94.9**

Table 6.6: Number of parallel sentences extracted from comparable corpora expanded by unidirectional and bidirectional approaches

Direction	No. of Extracted Pairs where Target Sentences are from:				Total Sentences	Name of Extracted Corpus
	Original Comparable Corpora	Machine-translated and Synthetic Datasets		Post-edited		
		NMT	mT5	Google Translate		
En→My	10,421	8,206	20,967	34,514	74,108	***s2t $\mathcal{P}$***
My→En	9,989	5,809	18,451	35,182	69,431	***t2s $\mathcal{P}$***
Bidirectional					132,857	***bd $\mathcal{P}$***

The proposed framework uses the bidirectional approach (i.e., source-to-target and target-to-source directions) for the Expand module for augmenting the target and source datasets that are used to expand the data size of existing comparable corpora. To compare the bidirectional approach with each unidirectional one (i.e., source-to-target or target-to-source), we implement both approaches in our experiments. The number of parallel sentences extracted by the Extract module from each approach (i.e., English-to-Myanmar unidirectional approach, Myanmar-to-English unidirectional approach, Bidirectional approach) are reported in Table 6.6. According to the report, it is clear that our bidirectional approach enables to expand more comparable data. The effectiveness of the extracted corpus is discussed in the next section.

6.5.4 Evaluation of Extracted Parallel Corpus on Machine Translation Task

We present the results in the context of a full machine translation system to evaluate the potential utility of the extracted parallel corpus. To evaluate the translation performance trained on existing and our extracted data, we use the BLEU score [Papineni et al., 2002] using the multi-bleu script from Moses[23]. In Table 6.7, we report the BLEU scores of Transformer-based NMT and mT5 systems on four different data size settings: only on existing parallel corpus, on existing corpus plus the ***s2t*** $\mathcal{P}$ corpus (parallel sentences extracted from comparable corpora after expanding the corpora with Myanmar sentences generated by the English-to-Myanmar forward-translation), on existing corpus plus the ***t2s*** $\mathcal{P}$ corpus (parallel sentences extracted from comparable corpora after expanding the corpora with English sentences generated by the Myanmar-to-English back-translation), and on existing corpus plus the ***bd*** $\mathcal{P}$ corpus (parallel sentences extracted from comparable corpora after expanding the corpora with all English and Myanmar sentences generated by both forward and back-translation).

Table 6.7: BLEU scores for English-Myanmar MT systems

Training Data	Total Sentences	English-to-Myanmar		Myanmar-to-English	
		NMT	mT5	NMT	mT5
Existing Parallel Corpus	204,535	8.11	13.49	11.59	15.49
+ ***s2t*** $\mathcal{P}$	+74,108	12.92	16.32	14.56	19.08
+ ***t2s*** $\mathcal{P}$	+69,431	12.75	16.01	14.06	18.93
+ ***bd*** $\mathcal{P}$	+132,857	**14.03**	**18.57**	**16.81**	**21.32**

Observe that both NMT and mT5 systems gain an increase on the performance with more additional data when training with either ***s2t*** $\mathcal{P}$ or ***t2s*** $\mathcal{P}$. While the ***bd*** $\mathcal{P}$ corpus is used as the additional training data, both NMT and mT5 systems achieve up to +6 BLEU scores in both translation directions. The translation accuracy of neural MT systems depends on both the size and quality of the training data [Zin et al., 2021]. Therefore, according to the findings, it implies that our proposed framework is an effective approach for creating a large and reliable parallel corpus in English-Myanmar low-resource setting.

6.5.5 Ablation Study and Analysis

Additionally, we conduct a qualitative analysis and ablation study on the effectiveness of the denoising-based automatic post-editing (DbAPE) system in this work. To validate the necessity of DbAPE, we further conduct the experiment without using it in our framework. Table 6.8 shows the number of parallel sentences extracted by our proposed framework without applying

[23] https://github.com/moses-smt/mosesdecoder

DbAPE in the data augmentation process of our Expand module. A comparison between Table 6.8 (i.e., without DbAPE) and Table 6.6 (i.e., with DbAPE) shows 23% (in English-to-Myanmar direction), 36% (in Myanmar-to-English direction), and 22% (in bi-direction) decrease on average on the number of extracted sentence pairs. From our analysis, it is because some of the forward and back-translated texts contain systematic errors such as extra words and missing words, and thus the source and translated sentences become less in similarity.

Table 6.8: Number of parallel sentences extracted from comparable corpora expanded by unidirectional and bidirectional approaches without using DbAPE system in the data augmentation processes.

Direction	No. of Extracted Pairs where Target Sentences are from:				Total Sentences	Name of Extracted Corpus
	Original Comparable Corpora	Machine-translated without Post-edited Synthetic Datasets				
		NMT	mT5	Google Translate		
En→My	10,421	6,319	19,447	20,618	56,805	***s2t*** $\hat{P}$
My→En	9,989	3,621	10,214	20,637	44,461	***t2s*** $\hat{P}$
Bidirectional					104,018	***bd*** $\hat{P}$

Table 6.9: BLEU scores for English-Myanmar MT systems. Without using DbAPE in data augmentation processes, it is not only decreased in the number of extracted sentence pairs but also in BLEU scores.

Training Data	Total Sentences	English-to-Myanmar		Myanmar-to-English	
		NMT	mT5	NMT	mT5
Existing Parallel Corpus	204,535	8.11	13.49	11.59	15.49
+ ***s2t*** $\hat{P}$	+56,805	11.87	15.62	13.07	17.36
+ ***t2s*** $\hat{P}$	+44,461	11.04	15.31	12.93	17.10
+ ***bd*** $\hat{P}$	+104,018	**13.86**	**17.93**	**15.74**	**20.63**

Table 6.9 shows the BLEU scores of the MT systems trained on additional data as mentioned in Table 6.8. The results show that DbAPE is a necessary component for the proposed framework to increase the amount of reliable parallel sentence pairs.

6.6 Summary

Our first substantial contribution is to demonstrate that collecting comparable monolingual texts to obtain a comparable corpus and expanding its size using the state-of-the-art data augmentation models is a useful approach for mining parallel data. The increasing volume of

extracted parallel sentences is a somewhat surprising result in improving low-resource MT systems. Hopefully this will encourage research into building large-scale comparable corpora for English-Myanmar low-resource machine translation.

Secondly, we improve on the candidate scoring function by effectively considering the usefulness of cosine similarity and margin-based scoring. This research points out that not only the cosine similarity but also the margin-based scoring alone is somehow less effective than their combination.

Moreover, initial investigations have shown that substantial gains can be achieved by applying the denoising-based automatic post-editing (DbAPE) model altogether with self-training and back-translation approaches. This enables to gain a higher number of high-quality similar the sentence pairs.

The extensive experiment shows that even with the small comparable corpora which contains small number of parallel sentences, our proposed framework can effectively expand the size of the comparable corpora and generate a little huge amount of reliable additional training data. Both NMT and mT5 systems can achieve up to +6 additional BLEU points in both translation directions for English-Myanmar language pair by using our created parallel corpus as the additional training data.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Our thesis is motivated by the fact that combining the effectiveness of optimized word segmentation, training data augmentation, and automatic post-editing for enhancing machine translation output will benefit for many low-resource language pairs. Deep learning and leveraging pre-trained word/sentence embeddings and monolingual data are a promising approach for solving that task.

The main contributions of this dissertation are summarized as follows:

- **Optimizing Myanmar Word Segmentation** (Chapter 3): We combine the advantages of NFKC normalization and byte-pair-encoding (BPE) mechanism via the proposed unsupervised Myanmar word segmenter. This approach efficiently solve out-of-vocabulary (OOV) problems in current MT systems. In addition, the proposed approach is a simple, cost effective and adaptable any MT domain at hand.

- **Denoising-based Automatic Post-editing** (Chapter 4): The proposed APE architecture, which employs pre-trained embedding model and denoising mechanism, achieves significant improvement and robustness in the post-editing task. We automatically carried out the post-editing task through three modules. The first modules is to find out mistranslation and missing information in machine-translated text (*mt*) via word alignment information. The second module enriched the *mt* by removing errors (mistranslations) and adding missing source size information. Then, the final module denoise the enriched version of *mt* to be in correct grammatical structure. For the first two modules, we simply design the workflow by applying pre-trained LaBSE model. For the final module, we use denoising autoencoder. In addition, the proposed architecture can be used as the post-processor not only for low-resource languages but also applicable for rich-resource language.

- **Building Parallel Corpus from Monolingual Target Texts** (Chapter 5): We augmented more parallel data for improving the performance of existing Myanmar-English MT systems. Leveraging pre-trained sentence embedding model to filter low-quality (less in similarity with huge semantic gap) sentence pairs from augmented corpus (pseudo-parallel corpus), consistently obtains high-quality training data. Through the experimental results, NMT achieved better performance using our extracted pseudo-parallel corpus as the additional training data.

- **Building Parallel Corpus from Comparable Corpora** (Chapter 6): We expanded our collected comparable corpora by augmenting more source and target sentences

using various MT approaches in order to obtain larger amount of reliable parallel data. We applied both self-training and back-translation approaches in our data augmentation process. Combining cosine similarity and margin-score function can extract reliable parallel sentence pairs from the expanded comparable corpora. Through the experimental results, both NMT and mT5 systems achieved better performance using our created parallel corpus as the additional training data.

7.2 Future Work

The next study will focus on the following things:

- **Optimizing Myanmar Word Segmentation** (Chapter 3): In our work, we just consider Myanmar word segmentation for machine translation task. It is interesting to apply our unsupervised approach to combination models (eg., name entity recognition (NER), syllable N-grams) for generating better word segmentation results.

- **Denoising-based Automatic Post-editing** (Chapter 4): Recently, reinforcement learning at informal-formal text style transfer achieves many successes. The combination of our APE architecture and reinforcement learning based text style transfer can lead to interesting results.

- **Building Parallel Corpus from Monolingual Target Texts** (Chapter 5): For future work, we plan to improve the performance of existing MT systems with more augmented data. Collecting comparable corpora, automatically extracting parallel corpus from collected corpora, and minimizing the semantic gap between extracted pairs by our proposed APE approach could lead to valuable results in high-quality training data creation task.

- **Building Parallel Corpus from Comparable Corpora** (Chapter 6): In the future, we would like to explore an alternative scoring method like a word alignment-based scoring approach. We would also like to explore strategies to exploit monolingual documents that are uploaded online in PDF format such as daily newspaper written in both English and Myanmar languages for building a large-scale comparable data. We will consider using the optical character recognition (OCR) tools for converting image to text and PDF file to Doc file, to easily copy the monolingual texts, and topic-level/document-level clustering approaches for aligning the copied English and Myanmar texts which are in the same clusters.

Bibliography

[1] K. Knight and I. Chander. Automated postediting of documents. In AAAI ’94, pp. 779–784. 1994

[2] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318, 2002.

[3] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference ofthe North American Chapter of the Association for Computational Linguistics, pages 127–133, 2003.

[4] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. Computational linguistics. 2003 Mar 1;29(1):19-51.

[5] C. Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pages 605–612, 2004a.

[6] C. Y. Lin and F. J. Och. Orange: A method for evaluating automatic evaluation metrics for machine translation. In Proceedings of the 20th International Conference on Computational Linguistics, pages 501–507, 2004b.

[7] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, 2006.

[8] D. Vilar, M. Popovi´c, and H. Ney, “AER: Do we need to “improve” our alignments?,” In Proceedings of the International Workshop on Spoken Language Translation, 2006.

[9] M. Simard, C. Goutte, and P. Isabelle, “Statistical Phrase-based Post-editing,” in Proc. NAACL HLT 2007:508–515, 2007a.

[10] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn, “Rule-based translation with statistical phrase-based Post-editing,” In: Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic, 2007b, pp 203–206.

[11] R. Zhang, K. Yasuda and E. Sumita. Chinese word segmentation and statistical machine translation. ACM Transactions on Speech and Language Processing (TSLP). 2008 May 29;5(2):1-9.

[12] Q. Gao and S. Vogel, “Parallel implementations of word alignment tool,” In Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pp. 49-57, Jun. 2008.

[13] M. Paul, E. Sumita, and K. Arora. Handling of out-of-vocabulary words in phrase-based statistical machine translation for Hindi-Japanese. Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing. March 2008.

[14] P. C. Chang, M. Galley, C. D. Manning. Optimizing Chinese word segmentation for machine translation performance. InProceedings of the third workshop on statistical machine translation 2008 Jun (pp. 224-232).

[15] D. Mareček, R. Rosa, P. Galuščáková , and O. Bojar. Two-step translation with grammatical post-processing. InProceedings of the Sixth Workshop on Statistical Machine Translation. 2011 Jul (pp. 426-432).

[16] H. Béchara, Y. Ma, and J. van Genabith, "Statistical post-editing for a statistical MT system," In: Proceedings of the 13th machine translation summit (MT Summit XIII), Xiamen, China, 2011, pp 308–315

[17] K. Parton, N. Habash, K. Mckeown, and G. Iglesias, Can automatic post-editing make MT more meaningful?. In: Proceedings of the 16th annual conference of the European Association for Machine Translation (EAMT 2012) pp 111–118. 2012.

[18] N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1700–1709. 2013

[19] A. Graves. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850, 2013.

[20] A. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2013.

[21] l. Sutskever, O. Vinyals, and Q. V. V. Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, pages 3104–3112, 2014.

[22] Y. K. Thu, A. Finch, E. Sumita, and Y. Sagisaka. Integrating dictionaries into an unsupervised model for Myanmar word segmentation. In Proc. of WSSANLP. 20–27. 2014.

[23] D. Bahdanau, K. H. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

[24] R. Sennrich, B. Haddow and A. Birch. Neural machine translation of rare words with sub-word units. arXiv preprint arXiv:1508.07909. 2015 Aug 31.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[26] H. Riza, M. Purwoadi, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, S. Sam, and S. Seng, "Introduction of the Asian language treebank," In 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) 2016 Oct 26 (pp. 1-6). IEEE.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016

[28] J. Zhang and C. Zong. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, 2016.

[29] M. Junczys-Dowmunt and R. Grundkiewicz, "Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing," in Proc. 1st Conf. Mach. Transl., Shared Task Papers, vol. 2, 2016, pp. 751–758.

[30] R. Sennrich, B. Haddow, and A. Birch. 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), pages 86–96, August 2016.

[31] P. Koehn and R. Knowles. Six challenges for neural machine translation. Proceedings of the First Workshop on Neural Machine Translation, abs/1706.03872:28–39. 2017.

[32] H. Schwenk, and M. Douze. Learning joint multilingual sentence representations with neural machine translation. arXiv preprint arXiv:1704.04154. 2017 Apr 13.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008. 2017

[34] J. Daems, S. Vandepitte, R. J. Hartsuiker, and L. Macken. Identifying the machine translation error types with the greatest impact on post-editing effort. Frontiers in psychology. 2017 Aug 2;8:1282.

[35] M. Fadaee, B. Arianna and M. Christof. Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440. 2017 May 1.

[36] A. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," CoRR, abs/1711.05101, 2017.

[37] F. Hieber, T. Domhan, M. Denkowski, D. Vilar, A. Sokolov, A. Clifton, and M. Post, "Sockeye: A toolkit for neural machine translation," arXiv preprint arXiv:1712.05690. 2017 Dec 15.

[38] M. Negri, M. Turchi, R. Chatterjee, and N. Bertoldi, "ESCAPE: A large scale synthetic corpus for automatic post-editing," in Proc. 11th Int. Conf. Lang. Resour. Eval., 2018, pp. 24–30.

[39] R. Chatterjee, M. Negri, R. Rubino, and M. Turchi, "Findings of the WMT 2018 shared task on automatic post-editing," in Proc. 3rd Conf. Mach.Transl., Shared Task Papers, 2018, pp. 723–738.

[40] Y. Kim, J. Geng, and H. Ney, "Improving Unsupervised Word-by-Word Translation Using Language Model and Denoising Autoencoder," EMNLP 2018.

[41] M. Guo, Q. Shen, Y. Yang, H. Ge, D. Cer, G. H. Abrego, K. Stevens, N. Constant, Y.

H. Sung, B. Strope, and R. Kurzweil. Effective parallel corpus mining using bilingual sentence embeddings. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 165–176. Association for Computational Linguistics. 2018.

[42] J. Zhang. Improving the transformer translation model with document-level context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 533–542, Brussels, Belgium, November 2018.

[43] H. Wang and Y. Lepage. Unsupervised Word Segmentation Using Minimum Description Length for Neural Machine Translation. Proceedings of the 24th Annual Meeting of the Association for Natural Language Processing. March 2018.

[44] N. F. Liu, J. May, M. Pust, and K. Knight. Augmenting statistical machine translation with subword translation of out-of-vocabulary words. arXiv preprint arXiv:1808.05700. 2018 Aug 16.

[45] M. Negri, M. Turchi, R. Chatterjee, and N. Bertoldi. ESCAPE: a large-scale synthetic corpus for automatic post-editing. arXiv preprint arXiv:1803.07274, 2018 Mar 20.

[46] X. Niu, M. Denkowski, and M. Carpuat, "Bi-directional neural machine translation with synthetic parallel data," *arXiv preprint arXiv:1805.11213*.

[47] G. Klein, Y. Kim, Y. Deng, V. Nguyen, J. Senellart, and A. M. Rush, "OpenNMT: Neural machine translation toolkit," arXiv preprint arXiv:1805.11462. May 28, 2018.

[48] Y. Yang, G. H. Ábrego, S. Yuan, M. Guo, Q. Shen, D. Cer, Y. H. Sung, B. Strope, and R. Kurzweil. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pages 5370–5378. ijcai.org. 2019

[49] A. Conneau and G. Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 7059–7069. Curran Associates, Inc. 2019

[50] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint. 2019.

[51] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116. Nov 5, 2019.

[52] R. Chatterjee, C. Federmann, M. Negri, and M. Turchi, ''Findings of the WMT 2019 shared task on automatic post-editing,'' in Proc. 4th Conf.Mach. Transl. (Shared Task Papers, Day), vol. 3, 2019, pp. 11–28.

[53] N. Reimers, and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. 2019 Aug 27.

[54] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep

bidirectional transformers for language understanding," In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019.

[55] A. V. Tetko, P. Karpov, R. V. Deursen, and G. Godin. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nature Communications, vol. 11, no. 1, pp. 1–11, 2020.

[56] T. Wolf. Transformers: state-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, Suzhou, China, November 2020.

[57] J. Ive, L. Specia, S. Szoc, T. Vanallemeersch, J. V. den Bogaert, E. Farah, C. Maroti, A. Ventura, and M. Khalilov, "A post-editing dataset in the legal domain: Do we underestimate neural machine translation quality?," In Proceedings of the 12th Language Resources and Evaluation Conference, 2020.

[58] T. Zenkel, J. Wuebker, and J. DeNero, "End-to-end neural word alignment outperforms GIZA++," arXiv preprint arXiv:2004.14675. Apr 30, 2020.

[59] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," arXiv preprint arXiv:2007.01852, Jul 3, 2020.

[60] W. Lee, J. Shin, B. Jung, J. Lee, and J.-H. Lee, ''Noising scheme for data augmentation in automatic post-editing,'' in Proc. 5th Conf. Mach. Transl., 2020, pp. 783–788.

[61] M. J. Sabet, P. Dufter, F. Yvon, and H. Sch¨utze, "SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings," arXiv preprint arXiv:2004.08728. Apr 18, 2020.

[62] F. do Carmo, D. Shterionov, J. Moorkens, J. Wagner, M. Hossari, E. Paquin, D. Schmidtke, D. Groves, and A. Way, "A review of the state-of-the-art in automatic post-editing", Machine Translation, 35(2), pp.101-143, Jun. 2021.

[63] W. Lee, B. Jung, J. Shin, and J. H. Lee. Adaptation of back-translation to automatic post-editing for synthetic data generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3685–3691, Apr 2021.

[64] Z. Y. Dou, and G. Neubig, "Word alignment by fine-tuning embeddings on parallel corpora," arXiv preprint arXiv:2101.08231. Jan 20, 2021.

[65] H. Moon, C. Park, S. Eo, J. Seo, and H. Lim, "An Empirical Study on Automatic Post Editing for Neural Machine Translation," IEEE Access. 2021 Sep 3;9:123754-63.

[66] W. Lee, B. Jung, J. Shin, and J. H. Lee, "RESHAPE: Reverse-Edited Synthetic Hypotheses for Automatic Post-Editing," IEEE Access. 2022 Feb 25;10:28274-82.

[67] A. Imankulova, T. Sato, and M. Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In Proceedings of the 4th Workshop on Asian Translation (WAT2017), pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

[68] W. Wang, I. Caswell, and C. Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1282–1292, Florence, Italy. Association for Computational Linguistics.

[69] Jaiswal, N., Patidar, M., Kumari, S., Patwardhan, M., Karande, S., Agarwal, P. and Vig, L., 2020, December. Improving NMT via Filtered Back Translation. In Proceedings of the 7th Workshop on Asian Translation (pp. 154-159).

[70] J. Smith, C. Quirk, and K. Toutanova, "Extracting parallel sentences from comparable corpora using document level alignment," in Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics, pp. 403-411, June 2010.

[71] A. Karimi, E. Ansari, and B. S. Bigham, "Extracting an English-Persian parallel corpus from comparable corpora," arXiv preprint arXiv:1711.00681, 2017.

[72] A. Azpeitia, T. Etchegoyhen, and E. M. Garcia, "Extracting parallel sentences from comparable corpora with STACC variants," in Proceedings of the 11th Workshop on Building and Using Comparable Corpora, pp. 48-52, May 2018.

[73] F. Gr´egoire and P. Langlais, "Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation," in Proceedings of the 27th International Conference on Computational Linguistics, pages 1442–1453, 2018.

[74] V. Hangya and A. Fraser, "Unsupervised parallel sentence extraction with parallel segment detection helps machine translation," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1224–1234, 2019.

[75] S. Steingrímsson, P. Lohar, H. Loftsson, and A. Way, "Effective bitext extraction from comparable corpora using a combination of three different approaches," in Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021), pp. 8-17, September 2021.

[76] M. J. Althobaiti, "A simple yet robust algorithm for automatic extraction of parallel sentences: a case study on Arabic-English Wikipedia articles," IEEE Access, 10, pp.401-420, 2021.

[77] G. Xu, Y. Ko, and J. Seo, "Improving neural machine translation by filtering synthetic parallel data," Entropy, 21(12), p.1213, 2019.

[78] P. J. Chen, J. Shen, M. Le, V. Chaudhary, A. El-Kishky, G. Wenzek, M. Ott, and M. A. Ranzato, "Facebook AI's WAT19 Myanmar-English translation task submission," arXiv preprint arXiv:1910.06848, 2019.

[79] M. M. Zin, T. Racharak, and N. M. Le, "Construct-Extract: an effective model for building bilingual corpus to tmprove English-Myanmar machine translation," in ICAART (2), pp. 333-342, 2021.

[80] M. M. Zin, T. Racharak, and N. M. Le, “DbAPE: denoising-based APE system for improving English-Myanmar NMT,” IEEE Access, June 2022.

[81] H. Khayrallah, and P. Koehn, “On the impact of various types of noise on neural machine translation,” in Proceedings of the 2nd Workshop on Neural Machine Translation andGeneration, pages 74–83, Melbourne, Australia. Association for Computational Linguistics. 2018.

[82] K. Wołk, E. Rejmund, and K. Marasek, “Multi-domain machine translation enhancements by parallel data extraction from comparable corpora” arXiv preprint arXiv:1603.06785, 2016.

[83] H. Afli, L. Barrault, and H. Schwenk, “Building and using multimodal comparable corpora for machine translation,” Natural Language Engineering, 22(4), pp.603-625, 2016.

[84] C. Chu, T. Nakazawa, and S. Kurohashi, “Integrated parallel sentence and fragment extraction from comparable corpora: a case study on Chinese--Japanese Wikipedia,” ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 15(2), pp.1-22, 2015.

[85] W. Ling, L. Marujo, C. Dyer, A. W. Black, and I. Trancoso, “Crowdsourcing high-quality parallel data extraction from Twitter,” in Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 426-436, June 2014.

[86] H. Schwenk, “Filtering and mining parallel data in a joint multilingual space,” arXiv preprint arXiv:1805.09822, 2018.

[87] M. Artetxe and H. Schwenk, “Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond,” Transactions of the Association for Computational Linguistics, 7, pp.597-610, 2019.

[88] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. AK R. Jobanputra, A. Sharma, S. Sahoo, H. Diddee, D. Kakwani, N. Kumar, and A. Pradeep, “Samanantar: the largest publicly available parallel corpora collection for 11 indic languages,” Transactions of the Association for Computational Linguistics, 10, pp.145-162, 2022.

[89] P. Lohar, D. Ganguly, H. Afli, A. Way, and G. J. Jones, “FaDA: fast document aligner using word embedding,” Prague Bulletin of Mathematical Linguistics, (106), pp.169-179, 2016.

[90] N. Ueffing, “Using monolingual source-language data to improve MT performance,” in Proceedings of the Third International Workshop on Spoken Language Translation: Papers, 2006.

[91] J. He, J. Gu, J. Shen, and M. A. Ranzato, “Revisiting self-training for neural sequence generation,” arXiv preprint arXiv:1909.13788, 2019.

[92] M. Artetxe and H. Schwenk, “Margin-based parallel corpus mining with multilingual sentence embeddings,” arXiv preprint arXiv:1811.01136, 2018.

[93] Y. K. Thu, W. P. Pa, M. Utiyama, A. Finch, and E. Sumita, “Introducing the Asian language treebank (ALT),” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1574-1578, May 2016.

[94] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, and C. Dyer, “Moses: open source toolkit for statistical machine translation,” in Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pp. 177-180, June 2007.

Publications and Awards

Journals

[1] M. M. Zin, T. Racharak, and N.M. Le. DbAPE: Denoising-based APE System for Improving English-Myanmar NMT. IEEE Access. 2022 Jun 22.

Conference Papers

[2] M. M. Zin, T. Racharak, and N.M. Le. Construct-Extract: An Effective Model for Building Bilingual Corpus to Improve English-Myanmar Machine Translation. InICAART (2) 2021 (pp. 333-342).

[3] M. M. Zin, T. Racharak, and N.M. Le. Expand-Extract: A Parallel Corpus Mining Framework from Comparable Corpora for English-Myanmar Machine Translation. ICTAI 2022.