Construct-Extract: An Effective Model for Building Bilingual Corpus to Improve English-Myanmar Machine Translation

May Myo Zin, Teeradaj Racharak and Nguyen Minh Le
School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan

Keywords: Neural Machine Translation, Myanmar Word Segmentation, Parallel Corpus Creation, Back-translation, Siamese-BERT Network.

Abstract: When dealing with low resource languages such as Myanmar, using additional pseudo parallel data for training machine translation systems is often an effective approach. As a pseudo parallel corpus is generated by back-translating target monolingual texts into the source language, it potentially contains a lot of noise including translation errors and weakly paired sentences and is thus required cleaning. In this paper, we propose a noisy parallel-sentences filtering system called *Construct-Extract* based on cosine similarity and Siamese BERT-Networks based cross-lingual sentence embeddings. The proposed system filters out noisy sentences by extracting high score sentence pairs from the constructed pseudo parallel data to finally obtain better synthetic parallel data. As part of the proposed system, we also introduce an unsupervised Myanmar sub-word segmenter to improve the quality of current English-Myanmar translation models that are potential to be used as backward systems for back-translation and often suffer from Myanmar word segmentation errors. Experiments show that the proposed Myanmar word segmentation could help the backward system to construct more accurate back-translated pseudo parallel data and using our extracted pseudo parallel corpus led to improve the performance of English-Myanmar translation systems in the two directions.

1 INTRODUCTION

Sentence-aligned parallel corpus is a prerequisite resource for building statistical and neural machine translation (SMT and NMT) systems. Generally, using large quantity parallel sentences to train a machine translation (MT) system enables to produce better translation results. However, for low-resource languages such as Myanmar, parallel corpora remain scarce due mainly to the cost of their creation. Findings in the literature show that there are two methods that support the construction of comprehensive parallel corpora. The first one is to extract nearly parallel sentence pairs from available topic-aligned parallel documents, called *comparable corpora* (Grégoire and Langlais, 2018; Hangya and Fraser, 2019). The second one is to use an automatic back-translation model trained on existing parallel data for creating new pseudo parallel corpus from the available target monolingual text (Xu et al., 2019).

It is worth noting that Myanmar is a resource-poor language and only small amount of English-Myanmar parallel sentence pairs are currently available to build baseline MT systems. Moreover, topic-aligned documents (i.e. comparable corpora) that contain an amount of semantically similar sentence pairs are not yet available. However, there are plenty of English monolingual data, which are in various domains and are accessible easily. Available monolingual English language data can be automatically backward translated into the Myanmar language for creating additional parallel corpus for training MT models.

This paper proposes the use of the target-side data (monolingual English sentences) throughout the back-translation approach for improving both source-to-target MT model (a.k.a. a forward model) and target-to-source MT model (a.k.a. a backward model). Indeed, we construct a pseudo parallel corpus and further extract only high quality sentence pairs from the constructed corpus. We apply the back-translation approach to construct English-Myanmar synthetic parallel data as a pseudo parallel corpus from collected in-domain English monolingual texts, in which the collected English monolingual sentences and existing training data are in the same domain. Then, the English-to-Myanmar MT model is used as a backward model. Apart from increasing the size of a corpus, word segmentation has been shown to be

Zin, M., Racharak, T. and Le, N.
Construct-Extract: An Effective Model for Building Bilingual Corpus to Improve English-Myanmar Machine Translation.
DOI: 10.5220/0010318903330342
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 2*, pages 333-342
ISBN: 978-989-758-484-8

helpful for improving translation tasks (Zhao et al., 2013). If the word segmentation step has many errors, a high accuracy translation task of the backward model may not be as expected. Existing Myanmar word segmentation tools produce massive rare words in MT tasks. In order to improve the performance of a backward model, we specifically propose an unsupervised Myanmar word segmentation approach based on the *NFKC*[1] normalization and byte pair encoding (BPE) (Sennrich et al., 2016b) mechanisms. The proposed segmentation approach can learn itself to adapt the current MT domain and significantly reduce the out-of-vocabulary (OOV) rate.

Although a back-translation approach can generate a large amount of synthetic parallel data, there is no guarantee of its quality. Data generated with the back translation might have noisy target translations (from monolingual data). Data quality plays an essential role in training both statistical and neural machine translation models. Especially, NMT models are very sensitive to noise in inputs. Therefore, using a constructed pseudo parallel corpus without filtering low-quality noisy sentences pairs may lead NMT systems to the performance degradation. Regarding synthetic data filtering, we propose a simple but effective approach called *Construct-Extract* that extracts only high-quality parallel sentence pairs from our constructed corpus. Our approach is based on the sentence-level cosine similarity of any two sentence vectors, i.e., vector representations of the back-translated synthetic source (Myanmar) sentence and the monolingual target (English) sentence. We calculate the sentence vectors on each sentence using Siamese BERT-Networks with an additional MEAN pooling layer.

The contribution of this paper is that we demonstrate the feasibility of improving performance on the Myanmar-English machine translation task by developing a neural-based bilingual corpus creation framework called *Contruct-Extract*. There are three important outcomes. First, we introduce a simple but effective unsupervised Myanmar word segmentation approach for improving the generated results of MT models that are potential to be used as back-translation models in pseudo parallel corpus construction. Second, we construct English-Myanmar pseudo parallel data from English monolingual texts by applying back-translation approach using improved English-to-Myanmar backward model. Third, we propose a Siamese BERT-Networks based approach to high-quality parallel sentences extraction (from our constructed corpus). Experiments on English-Myanmar translations demonstrate the efficacy of the proposed Myanmar word segmentation on improving current MT models that are potential to be used as backward systems in back-translation tasks, and our constructed-extracted pseudo parallel corpus on enhancing performance of the final MT models for bidirectional translation tasks.

[1]https://en.wikipedia.org/wiki/Unicode_equivalence.

2 CONSTRUCT-EXTRACT: A NEURAL-BASED FRAMEWORK FOR BUILDING BILINGUAL CORPUS

Our neural-based framework for building Myanmar-English bilingual corpus comprises two main modules for (1) pseudo parallel corpus construction and (2) high-quality parallel sentences pairs extraction. Figure 1 shows an overview of the system containing these components. Briefly, Figure 1 (a) depicts the first module containing the following two steps: improving the backward NMT system with the proposed Myanmar word segmentation and generating a pseudo parallel corpus through back-translation. Figure 1 (b) depicts the second module which uses the Siamese BERT-Network architecture for extracting high-quality sentence pairs from the corpus generated by the first module.

For Figure 1 (a), we construct more parallel translated texts through back-translation (Sennrich et al., 2016a) using a volunteer translator, i.e. an automatic back-translation of the 150k in-domain target monolingual English text into the source Myanmar language using pre-trained English-to-Myanmar backward 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. However, NMT still suffers from the out-of-vocabulary (OOV) issue due to the weakness of the current Myanmar word segmentation model. Hence, we also propose and apply a Myanmar word segmentation model for improving the performance of the backward NMT system. Our proposed Myanmar word segmentation model learns only on the current training data to segment the text that fits with the current MT domain. We train our segmentation model as follows:

- First, the model treats Myanmar sentences as raw streams of Unicode characters and normalizes them into canonical forms;
- Then, we apply the idea of byte pair encoding

(BPE) on normalized corpus to construct the appropriate vocabulary.

We explain each step of our proposed segmentation model in detail in Section 2.1.

Our constructed (back-translated) pseudo parallel corpus might have noisy target translations in Myanmar language. NMT suffers more sensitivity to noisy data compared to SMT. In some works, using additional back-translated data to train NMT will cause translation performance to deteriorate (Du and Way, 2017) or the performance will not be as good as expected. To investigate and overcome this problem, we present an extraction model that incorporates Siamese BERT networks with cosine similarity to filter only high quality sentence pairs. The whole extraction process is shown in Figure 1 (b), in which Siamese BERT network is applied to indicate similar sentence embeddings between sentence vectors $\boldsymbol{u}$, $\boldsymbol{v}$ 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.

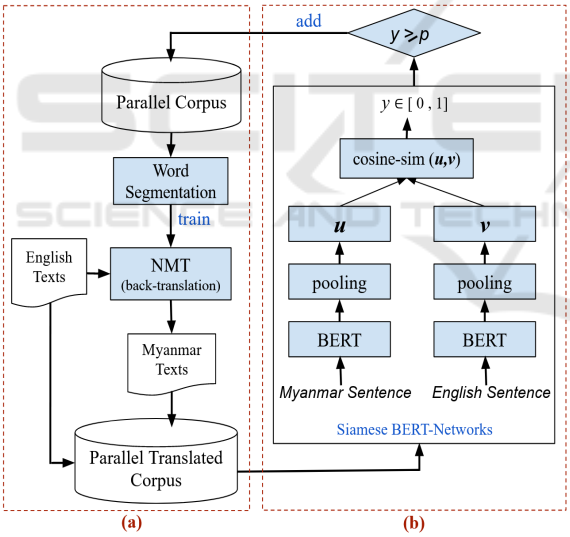

Figure 1: The proposed Construct-Extract framework for English-Myanmar parallel corpus creation.

2.1 Myanmar Word Segmentation

Our Myanmar word segmenter consists of three components: a normalizer, a trainer, and a tokenizer. Input sentences are treated as raw Unicode character streams, including the space as a use character. Figure 2 presents an overall architecture of the proposed Myanmar word segmenter.

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

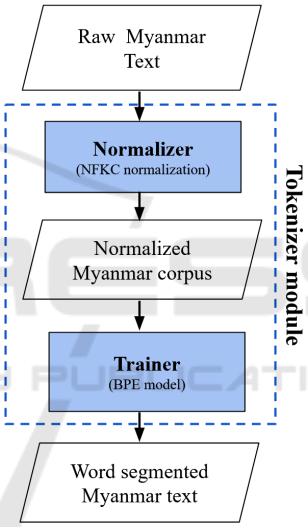

Figure 2: The proposed Myanmar word segmenter.

2.2 Back-translation

Back-translation approach (Sennrich et al., 2016a) is an effective data augmentation method leveraging target side monolingual data. We tried the back-translation to construct parallel translated sentence pairs from collected monolingual English texts. To perform back-translation, we first train English-to-Myanmar NMT (the backward system) with our proposed Myanmar word segmentation on the parallel data shown in Table 1, and use it to translate collected 150k English monolingual sentences to construct synthetic source side Myanmar texts. After the back-translation process, we constructed 150k English-Myanmar pseudo parallel sentence pairs. The noisy sentence pairs from the constructed corpus are then

removed with our proposed extraction module.

2.3 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. Currently, there are an increasing number of state-of-the-art pretrained models that support more than 100 languages including Myanmar and English. These models were trained based on the idea that a translated sentence should be mapped to the same location in the vector space as the original sentence. Therefore, they can generate aligned vector spaces, i.e., similar inputs in different languages are mapped closely in a vector space (Reimers and Gurevych, 2020).

In our experiments, we used the pre-trained model, i.e., *distilbert-multilingual-nli-stsb-quora-ranking* to derive semantically meaningful sentence embeddings between English sentences and back-translated Myanmar sentences. Then, we applied the cosine similarity to indicate how much an input sentence pair is semantically similar to each other. In our model, we threshold the similarity between each sentence pair at 0.77. If the similarity score between each sentence pair is greater than or equal to the threshold, the model decides to add that pair into the existing training data as a good quality parallel sentence pair. From 150 thousands parallel translated sentences, our model extracted only 92,111 sentence pairs. We examined the performance of SMT and NMT on existing datasets, with an additional 150k constructed dataset, and with an additional 92,111 extracted dataset to judge whether our proposed model can be effective or not. We elaborate this in more details in the next section.

3 EXPERIMENTS

This section describes the datasets and baseline MT systems that we have used in this work.

3.1 Datasets

We collected around 224 thousand manually created English-Myanmar parallel sentence pairs including bilingual sentences from text books, Myanmar local news, and the ALT Corpus (Riza et al., 2016) for training. The development and test sets are only from the ALT corpus. Data statistics are shown on Table 1. For the task of creating additional machine-translated pseudo parallel data, we additionally gather 150 thousands monolingual English sentences from the internet. These sentences nearly match the domain of the ALT corpus, which primarily contains news originated from English sources.

Table 1: Statistics of parallel datasets.

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

3.2 Baseline MT Systems

We evaluated the effectiveness of the proposed word segmentation model and the proposed Construct-Extract framework for construction and extraction of English-Myanmar parallel corpus by performing machine translation experiments.

3.2.1 Statistical Machine Translation

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

3.2.2 Neural Machine Translation

We trained the Transformer-based NMT models with PyTorch version of the OpenNMT project, 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 2.

Table 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

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, both of our proposed models are evaluated by performing statistical and neural MT systems: PBSMT and Transformer. Bilingual Evaluation UnderStudy (BLEU) score is used as the evaluation metric. The scores are computed using the multi-bleu script from Moses toolkit.

4.1 Effect of Word Segmentation

A lot of work has been done on the problem of Myanmar word segmentation and many word segmentation methods have been proposed. These segmentation methods can be roughly classified into dictionary-based or rule-based, statistical, machine learning and hybrid approaches (Pa and Thein, 2008; Ding et al., 2016; Phyu and Hashimoto, 2017; Oo and Soe, 2019). In the dictionary-based methods, only words that are stored in a pre-defined dictionary can be identified and the performance of the segmentation depends to a large degree upon the coverage of the dictionary. Increasing the size of the dictionary is not a good solution to the out of vocabulary word (OOV) problem because new words appear constantly. On the other hand, although the statistical approaches can somehow solve the problem of unknown words by utilizing probabilistic or cost-based scoring mechanisms, these methods also suffer from some drawbacks. The main issues are that they require large amounts of data for training, also with an amount of the processing time; and the difficulty in incorporating linguistic knowledge effectively into the segmentation process (Teahan et al., 2000). For low-resource languages such as Myanmar, there are only corpus-based, dictionary-based, rule-based, and statistical word segmentation methods freely available for being used as a temporary solution. Current Myanmar word segmentation tools can support to obtain better results for some Myanmar language processing tasks, such as part of speech (POS) tagging, word sense disambiguation, text categorization, information retrieval, text summarization, and etc. However, they may probably produce massive rare-words in both SMT and NMT. The segmentation error would cause translation mistakes directly especially in English-to-Myanmar translation. Although it is not a serious issue in Myanmar-to-English translation in general, weak Myanmar word segmentation tools can lead SMT to generate unknown source words as target translated words because they cannot find the corresponding target translation in the phrase table. The same problem also occurs in NMT.

Figure 3 illustrates some translation mistakes generated by the current English-to-Myanmar MT systems with currently available Myanmar word segmentation tools. These mistakes include: (i) missing words or phrases in the target Myanmar translation, (ii) translating English words into wrong Myanmar words, and (iii) generating both English words and their translated Myanmar words together in the translation results. Even in short sentence translation in a Myanmar-to-English direction as in Figure 4, it shows a problem of SMT that only generates an unknown Myanmar source word (in red) as a target English word and NMT misses to translate this word completely in translation. Note that this Myanmar source word (in red) should be translated as “fifteen” in the target English. This Myanmar word is formed by combining the two words (one word in green and another word in blue). The word (in green) is “fifteen” in English and the other word (in blue) is numerical classifier; it has no special meaning in English and is used only in Myanmar language that follows a number to show what type of thing that the number is referred to. This is because the current segmenter can only segment words based on their trained nature, corpus and dictionary that may not be fit with the available training corpus intended to use in MT tasks. In this case, Myanmar sentences are segmented using UCSYNLP word segmenter[2] that implements a combined model of bigram with word juncture and works by longest matching and bigram methods trained on a pre-segmented corpus of 50,000 words collected manually from Myanmar text books, newspapers, and journals (Pa and Thein, 2008).

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

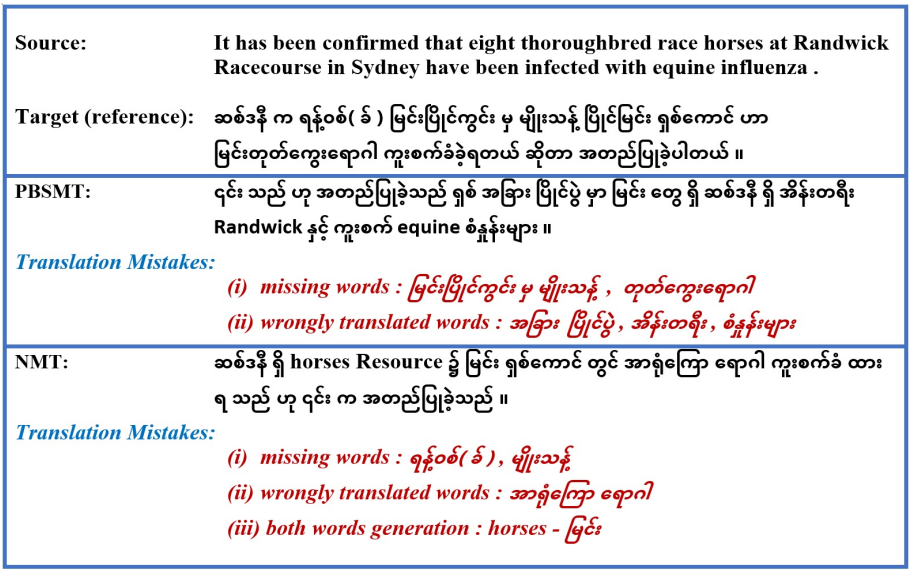

Figure 3: Translation errors of both statistical and neural English-to-Myanmar MT systems due to the Myanmar word segmentation weakness.

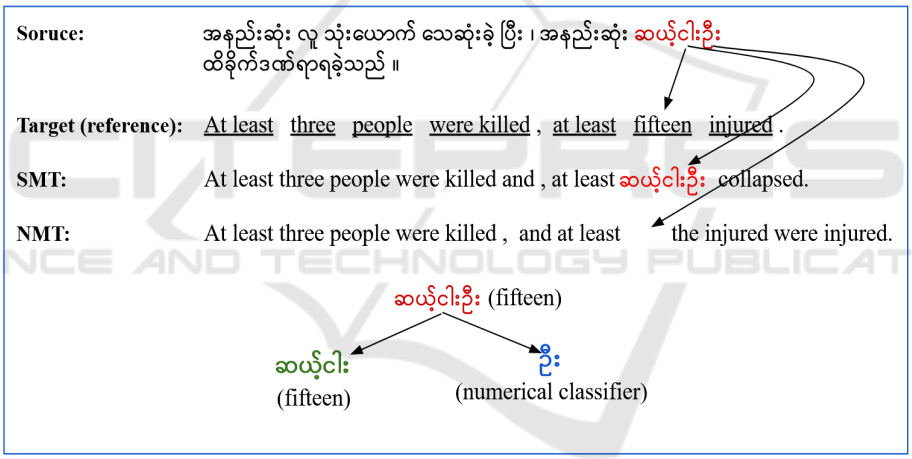

Figure 4: Translation errors of both statistical and neural Myanmar-to-English MT systems due to the Myanmar word segmentation weakness.

Our Myanmar word segmentation model does not require any linguistic resources and manual works. The only one requirement is to convert the text to segment into Unicode encoding. Currently, the Myanmar Unicode converters are freely available online and offline. The proposed model is able to learn on current MT corpus and thus it can produce the most suitable segmentation results. We analyse the effectiveness of our segmentation model in the MT experiments by checking the translated results and in terms of BLEU scores.

In our experiments, we used Moses tokenizer and truecaser for English texts. For Myanmar, UCSYNLP word segmenter is used as a baseline model. As explained in Subsection 2.1, our segmentation model consists of three components: a normalizer, a trainer, and a tokenizer. For our normalizer and trainer, we applied the same modules of Unicode *NFKC* Normalizer and BPE Trainer provided by SentencePiece (Kudo and Richardson, 2018). In the trainer process, we use a vocabulary size of 32,000 BPE sub-words. Our tokenizer module internally executes the normalizer to normalize the input Unicode character streams and tokenizes them into the word sequences with the segmentation model trained by the trainer.

Based on the proposed segmentation approach, the performance of MT systems are quite different. The results reported in Table 3 and Table 4, show-

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 မြို့တွင် 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 ရန့်ဝစ်(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 သူ့ကို မျှ ကျေ ကျေနပ် စုံစမ်း သေး ဟု ၊ လီ ဘ မန် က ပြောခဲ့သည် ။
Reference	**ဘယ် သတ်ဖြတ်မှုများ ကိုမျှ စိတ်ကျေနပ်ဖွယ် စုံစမ်းစစ်ဆေးခြင်း မပြုလုပ်ခဲ့ပါ ဟု ၊ ကော်ဗစ်လီ က ပြောကြားခဲ့သည် ။**

Figure 5: 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.

Table 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 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

ing that our unsupervised segmentation model can help the SMT and NMT systems to largely outperform the previous baselines. Our results have large gains on both MT systems in both directions. For the English-to-Myanmar task, SMT and NMT obtained a BLEU score of 7.63 and 8.11, respectively, with our proposed Myanmar word segmenter, which outperforms the previous best result by +3.48 and +2.86 points. For the Myanmar-to-English direction, NMT still surpasses the baseline score by 1.35 BLEU points. In this direction, the score of SMT is slightly decreased from 9.41 to 9.19. This is because we did not specifically care about names and numbers during the word segmentation process. Some of the rare names and numbers in the text are separated into two or three words, and is thus led to a little weakness only in the word alignment procedure of Myanmar-to-English PBSMT.

For investigating the OOV words issue, we used a copy mechanism in all experiments. The copy mechanism first tries to substitute OOV words with target words that have maximum attention weight according to their source words (Luong et al., 2015). When the words are not found, it copies the source words to the position of the not-found target words (Gu et al.,

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 6: A sample of constructed parallel sentences (monolingual English sentences and their corresponding back-translated Myanmar sentences). The Google translation of the back-translated Myanmar sentence in English is also provided. The score of how likely the sentences are semantically similar is calculated with cosine similarity. Only the first two sentence pairs that have similarity score of greater than 0.77 are extracted as the good quality sentence pairs.

2016). A detailed study of our results in English-Myanmar bi-directional translation tasks showed that the number of OOV words decreased considerably with our proposed Myanmar word segmentation. Figure 5 shows some example sentences generated by the English-to-Myanmar MT systems with the baseline (UCSYNLP segmenter) and with ours. Both SMT and NMT systems with our segmentation could handle the problem of OOV words than the outputs with baseline segmentation. The blue colored parts of the sentences in the figure are the correct translation parts in Myanmar language. It demonstrates that the NMT system leads to better translation accuracy and fluency than SMT. This part of experiments is done only on the existing parallel corpus. After obtaining the result, in all cases, we concluded that NMT with our segmenter is the best that can generate more accurate and fluent outputs. Therefore, for the next step of our back-translation in the parallel corpus construction task, we choose English-to-Myanmar NMT that was trained on existing corpus with our segmentation model as the volunteer pre-trained backward MT system.

4.2 Constructed-Extracted Data and Translation Results

The construct module of our proposed Construct-Extract model created 150k English-Myanmar pseudo parallel data by back-translating 150k monolingual English sentences into Myanmar language using English-to-Myanmar NMT (the backward system). After mixing these constructed corpus with the existing dataset, we have more training data to train on all MT systems from scratch. Generally, more training data help MT systems to improve the performance. However, there is one known challenge of NMT with low-quality noisy sentences. Some sentences in the back-translated corpus are low in quality. To investigate and overcome this challenge, we proposed the high-quality sentence pairs extraction module. Here, the extract module of our proposed model that are based on Siamese BERT-networks indicated only 92k high-quality parallel sentence pairs from the constructed corpus.

To evaluate the performance of our proposed approach, we manually looked at generated sentences and have done a qualitative analysis. In Figure 6, we can see the qualitative accuracy for some English-Myanmar parallel sentences constructed by using the back-translation approach. The back-translated Myanmar sentences have been translated into English using Google Translate, so as to facilitate a comparison with the original monolingual English sentences. Only the first two sentence pairs that have similarity score more than 0.77 are extracted as high-quality sentence pairs.

Table 5 and Table 6 illustrates the quality of the corpus created by the proposed Construct-Extract on

Table 5: 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 6: 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

the machine translation experiments. In these tables, we report 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). 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.

5 CONCLUSION

The motivation of this work is our expectation of improving the translation performance on the current English-Myanmar MT systems with the available limited resources. To set this goal, we present our two main contributions. The first one is Myanmar word segmentation model trained on the idea of Unicode *NFKC* normalization and the byte-pair-encoding mechanism. Our segmentation model is aimed to improve the performance of the backward system in pseudo parallel corpus construction task using a back-translation mechanism. The second one is a parallel corpus extraction methodology developed with the idea of Siamese-BERT-Networks-based sentence embedding and the cosine similarity. We validated the performance of these proposed models by performing SMT and NMT experiments.

Unlike traditional Myanmar segmenters that make use of manually prepared resources such as large-scale training data, dictionaries, etc, this proposed segmentation model does not need any manual work and any knowledge about Myanmar language. The model only requires converting Myanmar text written in other fonts into Unicode fonts with the use of freely available tools. Using our proposed segmenter on the preprocessing step of NMT systems, their translation performance improved quite a lot. On the other hand, the constructed and extracted parallel dataset is demonstrated to facilitate a significant improvement in MT quality when compared to a generic system as shown in our experimental results.

Overall, both of our Myanmar word segmenter and the parallel corpus extraction model are indeed beneficial for all MT systems to achieve a remarkable percentage which increases in the BLEU scores of the Myanmar-English low-resource problems, although the constructed corpus is less effective to support MT for yielding a significant BLEU score. We hypothesize that this is due to the lack of coverage on the sentence categories in the training and test datasets. More specifically, the training and test sets used by MT models in our experiments contains sentences from 13 different categories: crime and law, culture and entertainment, disasters and accidents, economy and business, education, the environment, health, obituaries, politics and conflicts, science and technology, sports, Wackynews, and weather. However, our constructed pseudo parallel corpus only covers 40 percent out of these categories. In the future, we plan to collect more monolingual corpus in different categories and extend

the proposed framework with a generative adversarial network for synthesizing high quality sentence candidates.

ACKNOWLEDGEMENTS

The authors would like to thank the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan for providing the Japanese Government (Monbukagakusho) Scholarship under which this work was carried out. This work was also supported in part by the Asian Office of Aerospace Research and Development (AOARD), Air Forced Office of Scientific Research (Grant no. FA2386-19-1-4041).

REFERENCES

Ding, C., Thu, Y. K., Utiyama, M., and Sumita, E. (2016). Word segmentation for Burmese (Myanmar). *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 15(4):1–10.

Du, J. and Way, A. (2017). Neural pre-translation for hybrid machine translation.

Grégoire, F. and Langlais, P. (2018). 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.

Gu, J., Lu, Z., Li, H., and Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Hangya, V. and Fraser, A. (2019). 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.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. M. (2018). Opennmt: Neural machine translation toolkit. *arXiv preprint arXiv:1805.11462*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Oo, Y. and Soe, K. M. (2019). Applying RNNs architecture by jointly learning segmentation and stemming for Myanmar language. In *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pages 391–393. IEEE.

Pa, W. P. and Thein, N. L. (2008). Myanmar word segmentation using hybrid approach. In *Proceedings of 6th International Conference on Computer Applications, Yangon, Myanmar*, pages 166–170.

Phyu, M. L. and Hashimoto, K. (2017). Burmese word segmentation with character clustering and CRFs. In *2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6. IEEE.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Riza, H., Purwoadi, M., Uliniansyah, T., Ti, A. A., Aljunied, S. M., Mai, L. C., Thang, V. T., Thai, N. P., Chea, V., Sam, S., et al. (2016). 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)*, pages 1–6. IEEE.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Sennrich, R., Haddow, B., and Birch, A. (2016a). 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, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Teahan, W. J., Wen, Y., McNab, R., and Witten, I. H. (2000). A compression-based algorithm for Chinese word segmentation. *Computational Linguistics*, 26(3):375–393.

Xu, G., Ko, Y., and Seo, J. (2019). Improving neural machine translation by filtering synthetic parallel data. *Entropy*, 21(12):1213.

Zhao, H., Utiyama, M., Sumita, E., and Lu, B.-L. (2013). An empirical study on word segmentation for Chinese machine translation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 248–263. Springer.