UCSYNLP-Lab Machine Translation Systems for WAT 2019

[1]Yi Mon Shwe Sin, [1]Win Pa Pa and [1]Khin Mar Soe
Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar
{yimonshwesin, winpapa, khinmarsoe}@ucsy.edu.mm

Abstract

This paper describes the UCSYNLP-Lab submission to WAT 2019 for Myanmar-English translation tasks in both direction. We have used the neural machine translation systems with attention model and utilized the UCSY-corpus and ALT corpus. In NMT with attention model, we use the word segmentation level as well as syllable segmentation level. Especially, we made the UCSY-corpus to be cleaned in WAT 2019. Therefore, the UCSY corpus for WAT 2019 is not identical to those used in WAT 2018. Experiments show that the translation systems can produce the substantial improvements.

1 Introduction

In recent years, Neural Machine Translation (NMT) (Bahdanau et al., 2015) as achieved state-of-the-art performance on various language pairs (Sennrich et al., 2016) and often outperforming traditional Statistical Machine Translation (SMT) techniques. Therefore, a lot of researchers have been attracted to investigate the machine translation based on neural methods. This paper describes the NMT systems of UCSYNLP-Lab for the WAT 2019 evaluation. We participated in Myanmar-English and English-Myanmar translations in both directions.

Although Myanmar sentences are clearly delimited by a sentence boundary maker but words or phrases are not always delimited by spaces. In Myanmar language, words are composed of one or more syllables and syllables are composed of characters. And syllables are not usually separated by white space. Therefore, word segmentation and syllable segmentation are essential steps for machine translation systems. Figure 1 describes the formation of Myanmar word and Myanmar syllable in one sentence.

English Sentence	The doctor gave me this prescription.																			
Myanmar Sentence	ဒီဆေးညွှန်းကဆရာဝန်ငါ့ကိုပေးလိုက်တာ။																			
Myanmar Phrases or clauses	Noun Phrase				Noun Phrase				Noun Phrase			Verb Phrase				Punctuation				
	ဒီဆေးညွှန်းက				ဆရာဝန်				ငါ့ကို			ပေးလိုက်တာ				။				
Myanmar Word	ဒီ		ဆေးညွှန်းက			ဆရာဝန်				ငါ့ကို			ပေးလိုက်တာ				။			
Myanmar Syllables	ဒီ	ဆေး	ညွှန်း	က	ဆ	ရာ	ဝန်		ငါ့	ကို		ပေး	လိုက်	တာ			။			

Figure 1: Formation of Myanmar sentence.

Moreover, Myanmar language is one of the low resource languages and there are a few parallel corpus. . It is necessary to be cleaned these corpus. So, we made the UCSY-corpus to be cleaned, therefore, the UCSY corpus for WAT 2019 is not identical to those used in WAT 2018. To enhance the performance of the model, we tried NMT with attention model with word level as well as syllable level. We employed NMT with attention model as our baseline model and built our translation system based on OpenNMT[1] open source toolkit.

The remainder of this paper is organized as follows: section 2 describes about the dataset. Section 3 describes the experimental set up and results are presented in section 4. Finally, we conclude in section 5.

2 Dataset

This section describes the dataset provided by WAT 2019 for the translation task. The datasets for Myanmar-English translation tasks at WAT2019 consists of parallel corpora from two different domains, namely, the ALT corpus and UCSY corpus. The ALT corpus is one part from

[1] http://github.com/OpenNMT/OpenNMT-py

Proceedings of the 6th Workshop on Asian Translation, pages 195–199

the Asian Language Treebank (ALT) project (Riza et al., 2016), consisting of twenty thousand Myanmar-English parallel sentences from the Wikinews. The UCSY corpus (Yi Mon Shwe Sin and Khin Mar Soe, 2018) is constructed by the NLP Lab, University of Computer Studies, Yangon (UCSY), Myanmar. The corpus consists of 200 thousand Myanmar-English parallel sentences collected from different domains, including news articles and textbooks.

ALT corpus size is extremely small, so a larger out-of-domain corpus for the same language pair also known as the UCSY corpus is provided. The UCSY corpus and a portion of the ALT corpus are used as training data, which are around 220,000 lines of sentences and phrases. The development and test data are from the ALT corpus. Therefore, the training data for Myanmar-English and English-Myanmar translation tasks is a mix domain data collected from different sources. Table 1 shows data statistics used for the experiments.

Data Type	File Name	Number of Sentences
TRAIN	train.ucsy.[my|en]	208,638
	train.alt.[my|en]	17,965
DEV	dev.alt.[my|en]	993
TEST	tet.alt.[my|en]	1.007

Table 1: Statistics of Datasets.

UCSY corpus was collected from bilingual sentences from various websites, and it contains some erroneous sentences, misspelled words, encoding problems and duplicate sentences. Therefore, we decided to remove these useless data after WAT 2018. Therefore, these problems are corrected manually at WAT2019 task to improve the quality of Machine Translation by removing duplicate sentences, spell checking, and normalizing different encodings.

3 Experimental Setup

We adopted a neural machine translation (NMT) with attention mechanism as a baseline system and we used OpenNMT[1] (Klein et al., 2017) as the implementation of the baseline NMT systems.

3.1 Training Data

The UCSY corpus and a portion of the ALT corpus are used as training data, which are around 220,000 lines of sentences and phrases. The development and test data are from the ALT corpus. Therefore, the training data for Myanmar-English and English-Myanmar translation tasks is a mix domain data collected from different sources. Table 2 shows the data about the training detail.

Domain	Number of Word		Myanmar Syllable tokens
	Myanmar	English	
ALT	698,347	436,923	1,138,297
UCSY	2,966,666	2,255,630	6,455,588
Total	36,650,13	2,692,553	6,569,417

Table 2: Training Details Information.

3.2 Tokenization

The collected raw sentences are not segmented correctly and some do not have almost no segmentation is essential for the quality improvement of Machine Translation. We used UCSYNLP word segmenter(Win Pa Pa and Ni Lar Thein, 2008) for Myanmar word segmentation and Myanmar syllable segmenter[2] for syllable segmentation.

UCSYNLP word segmenter is implemented a combined model, bigram and word juncture. This segmenter works by longest matching and bigram method with a pre-segmented corpus of 50,000 words collected manually from Myanmar Text Books, Newspapers, and Journals. The corpus is in Unicode encoding. After segementing the Myanmar sentence by UCSYNLP word segmenter the "_ " from the result is removed and replaced with space. Figure 2 shows the process of UCSYNLP word segmenter. It is not able to segment when "?" and "%" contains in Myanmar sentences. Examples are shown in Figure 3 and Figure 4. These sentences are segmented manually.

[2] https://github.com/ye-kyaw-thu/sylbreak

Before Segmentation	: အဲဒါကအဓိကပြဿနာပါ။
After Segmentation	: အဲဒါက_ အဓိက_ ပြဿနာ_ ပါ_ ။
Processing Step	: အဲဒါက အဓိက ပြဿနာ ပါ ။

Figure 2: The process of word level segmentation.

Before Segmentation	: ဟုတ်လား?ငါမသိလို့ပါ။
After Segmentation	: ဟုတ်လား_ ။ ငါ_ မသိ_ လို့ပါ_ ။

Figure 3: Sentences that are manually segmented.

Before Segmentation	: ကျောင်းသား၈၀%အောင်ပါတယ်။
After Segmentation	: Enter English Text

Figure 4: Sentences that are manually segmented.

For Myanmar syllable-based neural machine translation model, "sylbreak" is used to segment the Myanmar sentence into syllable level. Syllable segmentation is an important preprocess for many natural language processing (NLP) such as romanization, transliteration and grapheme-to-phoneme (g2p) conversion. "sylbreak" is a syllable segmentation tool for Myanmar language (Burmese) text encoded with Unicode (e.g. Myanmar3, Padauk). After segmenting the Myanmar sentence into syllable segmentation, the "|" from the result is removed and replaced with space and leading the trim process. Figure 5 shows the process of syllable segmentation for Myanmar syllable-based NMT model.

Before Segmentation	: တတ်ကြွလှပ်ရှားသူများကလည်းအလားတူစိုးရိမ်မှုများရှိတယ်။
After Segmentation	: ၊တတ်၊ကြွ၊လှပ်၊ရှား၊သူ၊များ၊က၊လည်း၊အ၊လား၊တူ၊စိုး၊ရိမ်၊မှု၊များ၊ရှိ၊တယ်။၊
Processing Step	: တတ် ကြွ လှပ် ရှား သူ များ က လည်း အ လား တူ စိုး ရိမ် မှု များ ရှိ တယ် ။

Figure 5: The process of syllable level segmentation.

3.3 NMT with attention

Our NMT system is built upon NMT with attention model that links blocks of Long Short-Term Memory (LSTM) in an RNN. We used open source OpenNMT. The experiments were run on Tesla K80 GPU. We trained the word-based NMT and Myanmar Syllable-based NMT. Based on different parameter settings, the training time is different. Table 3 shows the settings of network hyper-parameters for NMT models.

The basic architecture of the Encoder-Decoder model includes two recurrent neural networks (RNNs). A source recurrent neural network (RNN) encoder reads the source sentence x = (x1,…, xi) and encodes it into a sequence of hidden states h = (h1,…, hi). The target decoder is a recurrent neural network that generates a corresponding translation y = (y1,…, yj) based on the encoded sequence of hidden states h. The encoder and decoder are join to train to produce the maximum log-probability of the correct translation.

In attention based encoder-decoder architecture, encoder uses a bi-directional recurrent unit that gets a better performance for long sentences. Encoder encodes the annotation of each source word to summarize getting the preceding word and the following word. Likewise, the decoder also becomes a GRU and each word yj is predicted based on a recurrent hidden state, the previously predicted word yj-1, and a context vector. Unlike the previously encoder-decoder approach, the probability is conditioned on a distinct vector for each target word. This context vector is obtained from the weighted sum of the annotations hk, which is computed through an alignment model jk. Training is performed using stochastic gradient descent on a parallel corpus.

Hyper-parameter	NMT models
src vocab size	25,087 (Word Level)
tgt vocab size	50,004 (Word Level)
src vocab size	25,087(Syllable Level)
tgt vocab size	50,004 (Syllable Level)
Number of hidden units	500
Encoder layer	2
Decoder layer	2
Learning rate	1.0
Dropout rate	0.3
Mini-batch size	64

Table 3: Hyper-parameter of NMT models.

4 Experimental Results

Our systems are evaluated on the ALT test set using the evaluation metrics such as Bilingual Evaluation Understudy (BLEU) and Rank-based

Intuitive Bilingual Evaluation Score (RIBES). Table 4 and Table 5 show the different evaluation metrics for Myanmar-English and English-Myanmar translation pairs. We also investigated how segmentation level affects the MT performance in all experiments. The experimental results reveal that word level segmentation can give better performance for Myanmar to English NMT with attention model while syllable level segmentation can give better performance for English to Myanmar NMT.

	BLEU	RIBES
Word	**19.64**	**0.707789**
Syllable	15.96	0.657564

Table 4: Myanmar to English Translation.

	BLEU	RIBES
Word	14.84	0.697153
Syllable	**20.86**	**0.698507**

Table 5: : English to Myanmar Translation.

In Myanmar to English translation, word-based NMT model outperforms Myanmar Syllable-based NMT model in terms of BLEU score and the RIBES score. For Myanmar to English NMT system, word level segmentation NMT system performed much better than syllable level segmentation NMT system. That is, nearly 4 BLEU scores. However, Myanmar syllable-based NMT model gets higher score than word-based NMT in English to Myanmar translation. Interestingly, there is little difference in scores of RIBES in Myanmar syllable-based NMT model for English to Myanmar translation. For English to Myanmar NMT system, syllable level segmentation NMT system got the high BLEU scores that is nearly 6 BLEU scores. Best scores among those of the experimental results are submitted in this description.

5 Conclusions

In this system description for WAT2019, we submitted our NMT systems, which are NMT with attention. We evaluated our systems on Myanmar-English and English-Myanmar translations at WAT 2019. In the future, we will collect the more parallel sentences to get a large-sized MT corpus. And we also intend to do more and more experiments with more recent evolutions of the translation models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *In Proceedings of ACL – IJCNLP 2015*, Volume 1: Long Papers (2015). arXiv preprint arXiv:1409.0473.

Fabien Cromieres, Fabien Cromieres, Toshiaki Nakazawa and Toshiaki Nakazawa. Kyoto University Participation to WAT 2017, Proceedings of the 4th Workshop on Asian Translation, pages 146–153, Taipei, Taiwan, November 27, 2017. © 2017AFNLP.

Guillaume Klein, Yoon Kim, Yoon Kim, Jean Senellart, Alexander M. Rush, SYSTRAN and Harvard SEAS. OpenNMT: Open-Source Toolkit for Neural Machine Translation. (2017). Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 67–72 Vancouver, Canada, July 30- August 4, 2017. ©2017 Association for Computationsl Linguistics http://doi.org/10.18653/v1/P17-4012

Hammam Riza, Micheal Purwoadi, Gunarso, Tefuh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Rapid Sun, Vichet Chea, Khin Mar Soe, Khin Thandat Nwet, Masao Utiyama, Chenchen Ding, "Introduction of Asian Language Treebank with a Suvery of Asian NLP Resources", 2016.

Makoto Morishita, Jun Suzuki and Masaaki Nagata. NTT Neural Machine Translation Systems at WAT 2017. Proceedings of the 4th Workshop on Asian Translation, pages 89–94, Taipei, Taiwan, November 27, 2017. © 2017 AFNLP.

Minh-Thang Luong, Hieu Pham and Christopher D. Maiining. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natu-ral Language Processing, pages 1412-1421(2015).

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. IEEE Transactions on Audio, Speech, and Language Processing, 23(3):472-482.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 86–96.

Rico Sennrich, Barry Haddow and Alexandra Birch (2016): Neural Machine Translation of Rare Words

with Subword Units Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany

Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan and Pushpak Bhattacharyya. Comparing Recurrent and Convolutional Architectures for English-Hind Neural Machine Translation. Proceedings of the 4th Workshop on Asian Translation, pages 167–170, Taipei, Taiwan, November 27, 2017. ©2017 AFNLP .

Thet Thet Zin, Khin Mar Soe and Nilar Thein. Myanmar Phrases Translation Model with Morphological Analysis for Statistical Myanmar to English Translation System. 25th Pacific Asia Conference on Language, Information and Computation, pages 130-139(2011).

Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita, " A large scale study of Statistical Machine Translation Methods for Myanmar Language ", in Proc. Of SNLP2016, February 10-12, 2016.

Win Pa Pa, Ni Lar Thein. "Myanmar Word Segmentation using Hybrid Approach", Proceedings of 6th International Conference on Computer Applications, 2008, Yangon, pp-166-170.

Win Pa Pa, Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita. A Study of Statistical Machine Translation Methods for Under Resourced Languages. 29th Pacific Asia Conference on Language, Information and Computation pages 259-269(2016).

Yi Mon Shwe Sin and Khin Mar Soe, "Large Scale Myanmar to English Neural Machine Translation System". Proceeding of the IEEE 7th Global Conference on Consumer Electronic (GCCE 2018).