Bidirectional Neural Machine Translation for Myanmar-Korean-English Languages with Attention Mechanism

Yi Mon Shwe Sin[1], Hnin Nandar Zaw[2], Khin Mar Soe[1]

[1]Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar

[2]Faculty of Computer Science, University of Computer Studies, Pyay, Myanmar

yimonshwesin@ucsy.edu.mm

hninnandarzaw@ucsy.edu.mm

khinmarsoe@ucsy.edu.mm

***Abstract*— Nowadays, Neural Machine Translation (NMT) systems has been reached into a successful peak for rich-resource language pairs. However, NMT systems are hard to tackle to translate for poor-resource language pairs because of the insufficient number of parallel sentences. In this work, we present bidirectional Neural Machine Translation models with attention mechanism for three languages (Myanmar, Korean, and English). Although the amount of parallel data between these three language pairs are not very sufficient, this paper contributes the first evaluation of the neural machine translation models for Myanmar, Korean and English languages. And we also developed a Myanmar-Korean-English parallel corpus (around 37K sentences) based on the UCSY-corpus (Myanmar-English parallel sentences). Our experimental results are evaluated with Bilingual Evaluation Understudy (BLEU).**

***Keywords*— Neural Machine Translation, Myanmar, Korean, English, UCSY-corpus**

I. Introduction

Neural Machine Translation (NMT) system has shown the significant successful in some rich-resource languages in recent years. Unlike rule-based or statistical-based Machine Translation systems, Neural Machine Translation models learn to the end-to-end approach form the one language (source) to another language (target) directly. The essential component of Neural Machine Translation system includes an encoder, a decoder and an attention mechanism [1]. Although the performance and translation quality are continuously improvement, the translation results of the NMT models are highly depend on the size of parallel corpus for poor- resource language pairs. For this reason, creation of a parallel corpus for poor-resource language pairs becomes the primary challenge [2]. Therefore, data collection and data preparation are first step of our experiments. In order to create a Myanmar-Korean-English parallel corpus, some Myanmar sentences form the UCSY-corpus [6] are collected. Later, these Myanmar sentences are translated into Korean sentences manually by human experts who knows about the Korean Language [7]. More than 37K parallel sentences are created for the Myanmar-Korean-English parallel corpus.

In our work, our main motivation is to investigate Neural Machine Translation performance for Myanmar, Korean and English language pairs. In these days, some language pairs had successfully translated by developing neural methods. To train the proposed systems, the PyTorch OpenNMT tool [8] are used which are popular toolkit for natural language processing's researches.

In this paper, section 2 will describe related work of this paper. In section 3, the experimental settings of the system are expressed. Finally, the conclusions are presented in section 4.

II. Related Work

In [1], the authors first researched NMT with attention. Attention approach examine the mapping from the source words to the target words which is called a alignment model. This proposed model searches the most appropriate source information for each time and then predicts a target information according to this source information and all the previous generated target information. ACL WMT'14, English-to-French parallel corpus, is utilized. Two types of models, RNN Encoder-Decoder (RNNencdec) and the proposed model (RNNsearch), are trained with the same settings. It is showed that the proposed RNNsearch results is better than the conventional encoder–decoder model (RNNencdec).

In [3], the author presented the NMT systems with difference trainings, subword-level NMT training, training using the existing corpus plus monolingual data and a neural machine translation correction model for the Mongolian-Chinese language pairs. These three models are based on the attention-mechanism and CWMT'2009's Mongolian-Chinese parallel corpora (65K parallel sentences) are used. The system shows the proposed methods were improved the Mongolian-Chinese neural machine translation model.

III. Experimental setting

A. Dataset and Preprocessing

As the first step of our experiments, Myanmar-Korean-English parallel corpus are essentially created. Some Myanmar sentences from the UCSY-corpus [6] are collected.

UCSY-corpus is Myanmar-English parallel corpus which is a mix domain. From there, Myanmar sentences translated into Korean sentences manually by human experts who knows about the Korean Language [7]. More than 37K parallel sentences are created for the Myanmar-Korean-English parallel corpus which gives a good mix of short as well as long sentences. In the experiments, the training file, validation file and testing file are divided according to the Table 1.

Table 1. Data Statistics of parallel corpus

Files	**No. of sentences**
Training	33925
Validation	3017
Testing	700
Total Sentences	37642

At the preprocessing stage, word level tokenization is applied for each NMT training direction. UCSY NLP Word Segmenter tool [4] and Moses' [9] tokenization script was used for Myanmar sentences and English sentences respectively. And Korean sentences were segmented manually. Before the training, Moses's [9] clean script is also used for cleaning the file.

B. Models and Results

Nowadays, Neural Machine Translation systems have been succeeded in almost all languages especially for rich-resources language pairs. Additionally, there are numerous toolkits accessible for the research, development, and application of neural machine translation systems. There are various NMT implementations in use right now. The Myanmar-Korean-English neural machine translation models were developed using PyTorch OpenNMT's default settings, which is available on GitHub [8]. A vocabulary size of Myanmar language is 10,401 words, Korean language is 19,478 words and English language is 9107 respectively.

Table 2. Evaluation Result of the models

Source-Target	BLEU
MY - EN	16.24
EN- MY	22.19
KO - MY	20.32
MY- KO	12.58
KO-EN	21.42
EN-KO	19.86

The evaluation results of the experiments are shown in Table 2. MY, EN and KO are represented Myanmar Language, English Language and Korea Language respectively. It is found that the BLEU of MY-EN and MY-KO are less than another direction. Some foreign words cannot be translated into another direction. It is also found that most of the training data is only short sentences and few are long sentences. Therefore, the short sentences can be translated well in the translation results. Addition, the existing Myanmar-Korean-English parallel corpus are insufficient to train the translation models. To improve the quality of machine translation, the collection of more data and doing more works of other neural models are necessary to increase the translation performance.

IV. CONCLUSIONS

In this work, we present bidirectional Neural Machine Translation models for three languages (Myanmar, English, Korean) with six translation directions. And we also developed a Myanmar-English-Korean parallel corpus (around 37K sentences) based on the UCSY-corpus (Myanmar-English parallel sentences). Although the amount of parallel data between these three language pairs are not very sufficient, this paper contributes the first step of the machine translation models for Myanmar, English and Korean languages. To improve the quality of machine translation, the collection of more data and doing more works of other neural models are necessary as the future work.

REFERENCES

[1] D. Bahdanau, K. Cho, Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", Published as a conference paper at ICLR 2015.

[2] P. Koehn, R. Knowles, "Six Challenges for Neural Machine Translation", Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver, Canada, August 4, 2017.

[3] J. Wu, H. Hou, Z. Shen, J. Du, J. Li, "Adapting Attention-based Neural Network to Low-resource Mongolian-Chinese Machine Translation", (1777) ©Springer-Verlag Berlin Heidelberg 2011.

[4] W. P. Pa, N. Thein, "Myanmar Word Segmentation using Hybrid Approach", Proceedings of 6th International Conference on Computer Applications, 2008, Yangon, pp-166-170.

[5] Y. M. S. Sin, K. M. Soe, International Journal on Natural Language Computing (IJNLC) Vol.8, No.2, April 2019 "Attention-based syllable level neural machine translation system for Myanmar to English Language Pair"

[6] Y. M. S. Sin, K. M. Soe, K. Y. Htwe, "Large Scale Myanmar to English Neural Machine Translation System", Proceeding of the IEEE 7th Global Conference on Consumer Electronic (GCCE2018), Nara, Japan, 9th-12th October, 2018.

[7] H.N.D. Zaw, "Neural Machine Translation between Myanmar and Korean Languages", https://onlineresource.ucsy.edu.mm/handle/123456789/2775 , Dec, 2022,

[8] Pytorch-OpenNMT. http://github.com/OpenNMT/OpenNMT-py

[9] Moses Toolkit: http://www2.statmt.org/mo ses/