7th International Conference on Information, System and Convergence Applications
12 – 15 Feb. 2020, Ho Chi Minh, Vietnam

Data Augmentation Method for Korean Neural Machine Translation

Chanjun Park, Kinam Park, Heuiseok Lim
Korea University Dept. Computer Science, Seoul , Korea
bcj1210@naver.com, superkn@korea.ac.kr, limhseok@korea.ac.kr

Abstract— **Neural Machine Translation (NMT) has demonstrated much higher performance in the area of machine translation than any other system that is either rule-based or statistical. It is true that modeling plays an important role in Machine Translation. Nevertheless, this paper considers the idea that preprocessing is the most crucial step and thus performs various experiments regarding preprocessing. The subword tokenizer is the foremost step in machine translation. This paper demonstrates how different variations on implementing subword tokenization affects the quality of Ko-En machine translation. The tests on the Ko-En parallel corpus, which is open-source data, demonstrated that the best results occurred when the corpus was subword tokenized by the unigram-based Sentencepiece model after the data were separated on the basis of morpheme analysis. In addition, building large numbers of parallel corpora requires a great deal of time and money. To overcome these disadvantages, we can utilize a number of data augmentation techniques such as Back Translation and Copied Translation. When using Back Translation and Copied Translation, a monolingual corpus is converted into synthetic data to increase the amount of training data. This paper examines the effect of data augmentation on machine translation performance through various experiments that utilize Copied Translation as well as Back Translation. The results of the experiments showed that these data augmentation techniques helped improved machine translation performance. Furthermore, using a relative weighting ratio of original to synthetic data when constructing the batch of data significantly improved model performance.**

Keywords-component: Machine Translation, Back Translation, Copied Translation, Subword Tokenization,

I. INTRODUCTION

In the past, research on machine translation was performed by using rules and statistical based systems. But the current trend has shifted towards neural machine translation, resulting in a surge of public attention. [1, 2, 3] It is true that modeling plays an important role in Machine Transaltion. Nevertheless, this thesis considers preprocessing as the most crucial step and accordingly performs various experiments related to preprocessing. Subword tokenization is the foremost step in machine translation. The point of the preprocessing on the machine translation is how to subword tokenize each sentence. Generally, most theses make frequent use of the subword tokenization method using the BPE(Byte Pair Encoding) algorithm.[4] This paper proposes a subword tokenization method specialized in Korean-English machine translation by presenting several results from various tests.

Also, high-performance models such as the Transformer model have been trained with a large number of parallel corpora, and building parallel corpora is a time-consuming and expensive task. To overcome these shortcomings, there have been a number of data augmentations techniques such as Back Translation or Copied Translation that have been shown to resolve the issue of needing large amounts of training data. When using Back Translation and Copied Translation, a monolingual corpus is converted into synthetic data to increase the amount of training data. This paper examines the effect of data augmentation on machine translation performance through various experiments through Copied Translation as well as Back Translation.

II. SUBWORD TOKENIZATION

Subword tokenization is used to separate input sentences in machine translation into constant units. And this is the specific step of the preprocessing on MT to solve the out of vocabulary problem. It is quite common for other papers on the machine translation to tokenize with BPE[4].

A. Tokenization specialized in Korean

This paper suggests a more effective tokenization method than the one applied with only BPE throughout the multiple experiments. In Korean, due its nature as an agglutinative language, there is a postpositional particle that we refer to as josa. With this in mind, we have improved the performance on applying sentence piece unigram[5] by detaching josa from the input sentence after conducting morpheme analysis.

B. *Two Stage Subword Tokenization*

Three of the following methods are previously utilized methods of tokenization and the last two are newly created for the purposes of this paper:

(1) Byte Pair Encoding [4]

(2) SentencePiece Unigram Option [5]

(3) Morphological units

(4) Tokenize using Sentence Piece Unigram after "Josa" separation

(5) Tokenize using Sentence Piece Unigram after "Josa" separation + Compound Noun Decomposition

The 1st tokenization is generally used in the MT sphere and the 2nd one is well known as a more effective tokenization method that Google has implemented first in 2018. The 3rd method is tokenization based off of analyzing Korean morphemes. The 4th method is proposed by this paper as a method to separate josa based on the characteristics of Korean. It also aims to tokenize by sentence piece unigram after analyzing morphemes. The 5th tokenization method builds on the fourth tokenization method by implementing an additional feature that segments compound nouns.

III. Data Augmentation in Neural Machine Translation

There are two major data augmentation techniques in machine translation: Back Translation [6] and Copied Translation [7]. Back Translation is a method of translating a mono corpus using an existing trained reverse translator, creating a synthetic parallel corpus, and adding it to an existing bidirectional parallel corpus to use for training. It is based on the property that two translation models can be created with one parallel corpus. Copied Translation is a methodology that uses a mono-language corpus without the use of reverse translators.

IV. Experiments

A. *Data and Model*

As for the data, the open source Ko-En parallel corpus, called opensutitles2018 [1], was used in the Ko-En machine translation tests.
The model that was used for every experiment is the Transformer model [3]. The hyperparameters were the same hyperparameters as the ones that the model[3] is proposing. And the training was implemented on OpenNMT Pytorch[8]

[1] http://opus.nlpl.eu/OpenSubtitles-v2018.php

B. *Subword Tokenization Results*

TABLE I. Results of Subword Tokenization Experiments

	BLEU
Byte Pair Encoding	9.78
SentencePiece	10.66
Morphological Unit	11.79
SentencePiece+Morphological Unit	12.53
SentencePiece+Morphological Unit + Compound Noun Decomposition	12.03

Of the previously utilized tokenization methods such as BPE or SentencePiece, the morpheme based tokenization method results in the highest score. However, when utilizing the proposed tokenization model that first splits josa from each sentence and then applies the unigram-based SentencePiece model, the resulting BLEU score is significantly higher than the previously mentioned three existing methods of tokenization. Interestingly, when compound noun decomposition was added to the proposed model, the BLEU score remained higher than the other three existing models but proved to be lower than the proposed model without the compound noun decomposition. This implies that if the corpus is separated into units that are too small, this may have a negative impact on the performance of the tokenization model.

C. *Data Augmentation Results*

Now that we have determined that the proposed tokenization model results in the highest BLEU score, we can turn to finding the most effective method of data augmentation. BLEU score was used as an indicator of performance evaluation:

TABLE II. Experimental Results of Data Augmentation Experiments

	BLEU
BASE	12.53
Back Translation	12.89
Copied Translation	9.81
(Back+Copied) Translation	12.62

In the case of BASE, the model was trained without using any data augmentation techniques. Experimental results show that the model using Back Translation and Copied Translation show a higher BLEU score than the existing BASE; however, performance is much lower

when using only Copied Translation. This suggests that Copied Translation can be used only in conjunction with Back Translation. This is because there are a few words that are initially shared because the character sets of language pairs are different.

TABLE III. EXPERIMENTAL RESULTS OF APPLYING RELATIVE RATIO BETWEEN ORIGINAL CORPUS AND SYNTHETIC CORPUS

	BLEU
Back Translation 2:1 Ratio	13.22
Back Translation 3:1 Ratio	13.01
Back Translation 4:1 Ratio	12.96
Back Translation 3:2 Ratio	12.92
Back Translation 4:3 Ratio	12.98
(Back+Copied) Translation 2:1 Ratio	13.06
(Back+Copied) Translation 3:1 Ratio	13.09
(Back+Copied) Translation 4:1 Ratio	12.86
(Back+Copied) Translation 3:2 Ratio	12.77
(Back+Copied) Translation 4:3 Ratio	13.29

In the experiments above, we tested utilizing various ratios of original to synthetic corpora when constructing a batch to demonstrate the effects of various ratios on the BLEU scores.
All experiments were performed when the ratio was either 2 to 1, 3 to 1, 4 to 1, 3 to 2, or 4 to 3.
In the experiment, Back Translation and Copied Translation were applied together and we found that applying the relative ratio of 4 to 3 resulted in the highest BLEU score. This suggests that a large number of synthetic corpora does not necessarily always create a high performance model, and that training with a reasonable ratio of original corpora to synthetic corpora can also produce a model with good performance. In conclusion, data augmentation techniques such as Back Translation and Copied Translation help to improve machine translation performance.

V. CONCLUSION

There exist researchers working on the model itself in the research field on NMT. However, this paper suggests that the initial step of preprocessing is just as important as well. Further research on the preprocessing methods of not only Korean but also English and other European languages are scheduled to be done in the near future. This paper has also conducted various experiments on data augmentation techniques for machine translation. Data augmentation techniques such as Back Translation and Copied Translation help improve the performance of machine translation, and the experiments found that placing a relative weight when constructing a batch helps the performance as well. In the future, the parallel corpus filtering technique will be applied to the synthetic corpus to improve the machine translation performance.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks”. Put sponsor acknowledgments in the unnum-bered footnote on the first page.

REFERENCES

[1]Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation By Jointly Learning To Align and Translate. In ICLR, pages 1–15
[2]Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proc of EMNLP.
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.
[4] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proc. of ACL
[5] Taku Kudo, John Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, EMNLP2018
[6] Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving neural machine translation models with monolingual data." arXiv preprint arXiv:1511.06709 (2015).
[7]Edunov, Sergey, et al. "Understanding back-translation at scale." arXiv preprint arXiv:1808.09381 (2018).
[8] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. CoRR, abs/1701.02810.