The 3rd International Conference on Interdisciplinary research on
Computer science, Psychology, and Education (ICICPE' 2019)
December 17-19, 2019. Phu Quoc, Vietnam

Improved Machine Translation Performance Using Two-Stage Subword Tokenization

Chanjun Park[1], Sora Choi, Heuiseok Lim[1]
[1] Korea University, Computer Science
bcj1210@naver.com, srchoiforlove@gmail.com, limhseok@korea.ac.kr

Abstract. Neural Machine Translation (NMT) using Deep-learning has been showing much greater performances on the area of the machine translation than any other systems based on the rule-based or statistical ones. It is true that the modeling plays an important role in MT. Nevertheless, this thesis considers the preprocessing the most crucial step and thus performs various experiments related to it. The subword tokenizer is the foremost step in the machine translation. This thesis presents the main facts of subword tokenize reinforcing the quality of Ko-En machine translation through the multi experiments. As a result of the tests on the Ko-En parallel corpus which is an open source data, it was the best score when the corpus were subword tokenized by sentence piece model of which option is unigram based after the data were separated in a basis of morpheme analysis.

Keywords: Subword Tokenization, Machine Translation, Transformer

1 Introduction

In the past, the research on the machine translation was performed by using rules and statistical based systems. But the state of the art trend was changed into Neural machine translation and it has brought a broad of attention publicly. [1,2,3] The point of the preprocessing on the machine translation is how to subword tokenize each sentence. Generally, most theses make frequent use of subword tokenize method using BPE(Byte Pair Encoding) algorithm.[4] This paper proposes subword tokenize method specialized in the Korean-English machine translation by presenting the several results on tests.

2 Tokenization specialized in Korean

The tokenization is to separate input sentences in the machine translation into the constant unit. And this is the specific step of the preprocessing on MT to solve the out of vocabulary. It is quite common for lots of other papers on the machine translation to tokenize with BPE. This paper suggests much greater tokenization method than the one applied with only BPE throughout the multiple experiments. There is josa in Korean which is capable of being expressed in several ways as an inflectional language. Based on this inspirational fact, the performance on applying sentence piece unigram[5] by detaching josa after analyzing morpheme has been improved.

3 Experiments

3.1 Data & Model

As for the data, the open source of Ko-En parallel corpus, which is English subtitle called opensutitles2018[1], was used to have Ko-En machine translation test.

The model that was used for every experiment is Transformer model [3] and hyperparameters are as follows. The hyperparameters were used as all the same ones that the model[3] is proposing. And the training was implemented on the OpenNMT Pytorch[6]

3.2 Two Stage Subword Tokenization

3 methods will be seen as previous ones for the test and 2 methods are newly suggested.

1. Byte Pair Encoding
2. SentencePiece Unigram Option
3. Morphological units
4. Tokenize using Sentence Piece Unigram after "Josa" separation
5. Compound Noun Decomposition + "Josa" Separation + Tokenize Using SentencePiece Unigram Option

The 1st tokenization is generally used in the MT sphere and the 2nd one is well known as better tokenization method that google made a suggestion in 2018 than the first one. The 3rd one is the tokenization in a basis of analyzing Korean morpheme. The 4th one is proposed by this thesis as a method to separate josa based on the characteristics of Korean. It also aims to tokenize by sentence piece unigram after analyzing morpheme. The 5th one is to add the functions of segmenting compound nouns to the 4th method. The segmenter of compound words uses the one of [7]. The results are as follows.

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

3.3 Results

	BLEU
BASE	7.08
BPE	9.67
SentencePiece Unigram	10.49
MECAB	11.57
Mecab + SentencePiece Unigram	**12.22**
Mecab + SentencePiece Unigram + Compound Noun Decomposition	11.79

The base model doesn't consider tokenization but does filter parallel corpus. At the level of existing tests which are BPE, sentence piece unigram and tokenization in a basis of morpheme, the morpheme based tokenization gets the best score on the test among them. That is, to build the MT related to Korean, the morpheme based tokenization has validated the better quality than tokenization using BPE or sentence piece. It is shown that the BLEU score from the proposed tokenization using sentence piece unigram after splitting josa in each sentence through the morpheme analyzer has improved compared to three other existing methods. When compound nouns were segmented additionally and its model presented higher score than three other existing methods. But the score was lower than the model that used sentence unigram after josa segmented with the morpheme analyzer. It implies that if the corpus are separated into the smaller units, it may rather have a negative impact on the performance.

4 Conclusion

There exist researches working on the model itself in the research field on NMT. However it is considered in this paper that the preprocessing is important as well. The further researches on the preprocessing methods of not only Korean but English, other European languages and etc. are scheduled to be done in the near future.

5 Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2017M3C4A7068189).

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. 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.
7. Chanjun Park, Pummo Ryu 2018, "Two Stage Korea Compound Noun Decomposer", HCLT2018, pp. 495-497