Toward High Quality Parallel Corpus Using Monolingual Corpus

Chanjun Park[1], Youngdae Oh[2], Jongkeun Choi[3] , Dongpil Kim[4] , Heuiseok Lim[5*]
[1] *Master&Ph.D Combined Student, Department of Computer Science and Engineering*
Korea University, Seoul
[2]*Director, LLsoLLu, Seoul*
[3]*General Manager, LLsoLLu, Seoul*
[4]*Senior Executive Vice President, LLsoLLu, Seoul*
[5*] *Professor, Department of Computer Science and Engineering Korea University, Seoul*

bcj1210@korea.ac.kr[1], youngdae.oh@llsollu.com[2], jongkeun.choi@llsollu.com[3],
dongpil.kim@llsollu.com[4], limhseok@korea.ac.kr[5]*
Corresponding author: Heuiseok Lim*

Abstract It is very important to build high quality training data in deep learning. However, in machine translation, getting high quality parallel corpus is not easy due to copyright issues and the alignment. Also building a parallel corpus takes a lot of time and money, and most people only have mono corpus through web crawling. For these people, we propose an efficient high-quality parallel corpus construction process that can coexist with the human translation market.

Keywords*: Machine Translation, Parallel Corpus, Human Translation, High Quality Data, Deep Learning, Pseudo Parallel Corpus*

1. Introduction

Neural Machine Translation (NMT) open source has been activated, and many companies around the world are trying to develop and commercialize it. However there are insufficient data resources to achieve performance enhancement and advancement of the NMT solution. The most important thing in Deep Learning(DL) is to construct high quality training data. In other words, it is important to secure high-quality parallel corpus when building NMT model.

However, obtaining the parallel corpus is not easy due to the problem of securing copyright, difficulties in alignment, and numerous noises. In order to train the NMT model, a parallel corpus which contains at least 2 million lines is required, but it is not easy to prepare such a large corpus for training.

Therefore, this paper suggests the construction direction of high-quality parallel corpus and suggests a data construction process that can coexist with the human translation market. In other words, we propose a methodology for constructing a high-quality parallel corpus by using only mono corpus. Through this, it is expected to save a lot of time and money, and it is a process that enables human translation and the NMT market to coexist.

2. Data Construction Directionality

The proposed construction direction is largely divided into four: domain-based parallel corpus construction, balance corpus construction, parallel corpus filtering and refining, and human translation.

Documents can be classified into various domains. However, it is difficult for NMT model to accurately translate all fields (domains). When constructing data based on domain, it is possible to select and provide necessary data to users, and is easy to manage.

The balance corpus means a corpus composed of precise alignment and translation, and a corpus that contains various domains comprehensively. In other words, it is important to build a parallel corpus which includes various domains comprehensively, not to build a parallel corpus with only one domain.

There are a number of studies that a model trained with a corpus which has been refined and filtered has a higher BLEU score than a model that does not[1]. A model trained with a corpus selected through parallel corpus filtering shows a better BLEU score than a model which does not.

In order to improve the quality of data, it is the most reliable and high-quality data that can ultimately be put through human hands. However, if we build the data through human hands for all data, we will have to invest a lot of money and time.

Therefore, if the computer automatically judges the quality to a certain extent, and if the quality is above a certain level, it would be better to perform verification and post−processing after only the data below a certain level without the human hand.

3. The Proposed Method

Building parallel corpus is time consuming and waste of money, and most people only have a mono corpus through web crawling. For those people, we propose a process to produce high quality parallel corpus using only mono corpus. The overall process is shown in Figure 1.

First, the quality of mono corpus is improved by using mono corpus cleaning and grammar error correction. Because the data obtained through web crawling is often wrong in grammar and is unverified data. Subsequently, the Pseudo Parallel Corpus(PPC) is created using a pre−built Neural Machine Translation (NMT) system. After that, post processing is performed using the Automatic Post Editing(APE)[2] model for the PPC which has been constructed so far. Through this, the quality of PPC can be further improved. Subsequently, quality estimation[3] is performed by inputting the source sentence and the translation result after APE. Pearson's correlation, Mean Average Error (MAE), and Root Mean Squared Error (RMSE) are used for the sentence−level performance evaluation, and multiplication of F1−OK, F1−BAD is used for the performance evaluation for the word and phrase level. Using this score, the level of editing by human translation is determined[3]. It is classified into a total of 3 levels (High, Middle, and Low). High will be used as training data as it is. The remaining Middle and Low apply the Back Translation[4] to conduct the 2nd verification. After that, it is decided whether or not to proceed with the human translation process for the data. Through this, it is expected that high−quality parallel corpus will be built through the second verification process of professional translators. In addition, when entrusting the verification, different prices for Middle and Low should be measured.

The advantage of this process is that the cost of human translation can be reduced by setting the price differently. In the case of high level, since it is already high quality, it is easy to carry out the supervision work even if do not perform the editing work or take a little time. On the other hand, in the case of the low level, intensive verification is required. This can shorten the time and improve the efficiency of the verification work. In conclusion, it is possible to secure a high−quality parallel corpus by saving time and money and improving the efficiency of the verification work.

4. Conclusion

This paper presents the direction and process of building a high quality parallel corpus, one of the important elements in NMT. In the future, we will build a parallel corpus based on the process proposed in this paper.

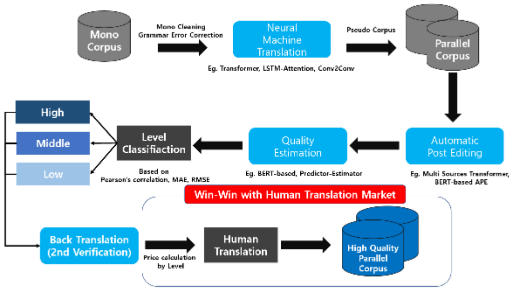
Mono Corpus
Mono Cleaning
Grammar Error Correction
Neural Machine Translation
Pseudo Corpus
Parallel Corpus
Eg. Transformer, LSTM-Attention, Conv2Conv
High
Middle
Low
Level Classifiaction
Quality Estimation
Automatic Post Editing
Based on Pearson's correlation, MAE, RMSE
Eg. BERT-based, Predictor-Estimator
Eg. Multi Sources Transformer, BERT-based APE
Win-Win with Human Translation Market
Back Translation (2nd Verification)
Price calculation by Level
Human Translation
High Quality Parallel Corpus

Figure 1. Process to build high quality Parallel Corpus

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2017M3C4A7068189)

References

[1] Koehn, Philipp, et al. "Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions." Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). 2019.

[2] Chatterjee, Rajen, et al. "Findings of the WMT 2019 Shared Task on Automatic Post-Editing." Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). 2019.

[3] Fonseca, Erick, et al. "Findings of the WMT 2019 Shared Tasks on Quality Estimation." Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). 2019.

[4] Edunov, Sergey, et al. "Understanding back-translation at scale." arXiv preprint arXiv:1808.09381 (2018).