The 4th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2020)
December 21-23, 2020. Jeju Island, Korea. (Online)

Domain-specialize Neural Machine Translation Methodology

Chanjun Park[1], Heuiseok Lim[1]

[1] Korea University, Computer Science
bcj1210@naver.com, limhseok@korea.ac.kr

Abstract. A commercialize Neural Machine Translatioon (NMT) model such as Google Translator is being serviced as one general model covering various damains. However, in the case of special domains such as COVID-19, patent, thesis, security, bio, etc., there are many limitations in translating to a general NMT model. Translating terminology or new words used only in a specific domain is because there are many difficulties in translating the general model. Accordingly, research on domain-specialize machine translation is underway, and training is conducted with data specialized in a specific domain, so that the quality of translation in that domain is superior to the general model. In line with these research trends, this paper proposes various methodologies for training domain-specialize NMT, and aims to present standards for them.

Keywords: Domain Specialized Machine Translation, Neural Machine Translation, Machine Translation, Corpus Weight Training, Incremental Training

1 Introduction

The problems of the latest neural machine translations[1,2,3] can be divided into three major problems: slow speed, extreme performance differences according to domains, and lack of high-quality and large-scale data. In the first case, it is a problem caused by applying a finetuning approach based on a pretrain model such as BERT[4] or MASS[5], which causes the disadvantage that the model size is too large and the decoding speed is slow.

In the second case, in general, most companies provide services based on one general model, and accordingly, there is a problem that there is a significant difference in translation quality depending on the domain. Therefore, research is needed to separately construct a domain-specific model.

Third, it is difficult to obtain high-quality, large-scale data, which is one of the important factors in deep learning, and furthermore, it is more difficult to construct domain-specific data. This is because a lot of time and money must be invested in building domain-specific data.

Therefore, in order to make money by conducting business using machine translation, there is no choice but to create a translator specialized for the domain. This is because the commercialized system is open for free for general machine translators, and there are limitations in translating patent documents and thesis with this model. Therefore, it is considered that machine translation research that ultimately generates domain-specific sentences will develop into one mainstream research, and it is expected that this research will be considerably needed by companies and research institutes in the future. Therefore, this paper intends to standardize by defining four methodologies that can train domain-specific machine translators.

2 Domain Specialized Neural Machine Translation Methodologies

There is a need to proceed with research to find an optimal specialized methodology after establishing various strategies to make a domain-specific machine translator. Therefore, this paper intends to present four strategies.

First is Incremental Training / Re-training methodology. This methodology is a simple method to proceed with the Fine Tuning Approach to domain data based on the General model. In addition, it is also possible to perform performance comparison experiments according to the increasing amount of data through incremental training. This will be a form similar to Curriculum Learning. Figure 1 shows a architecture of this methodology.

The second is Domain Ensemble Decoding. It is possible to check whether the performance is improved through ensemble decoding of the basic model and domain-specific model.

Third is Combine Augmentation. Performance comparison with Incremental Training is possible through Combine Augmentation for general data and domain data. Combine Augmentation is a methodology in which domain data and general data are simply combined and trained.

Fourth is General-Domain Corpus Weight Training. The study can be conducted by applying the Corpus Weight methodology, which makes the ratio of the general corpus and the domain corpus relatively. That is, when construct a batch, it is a methodology that establishes a relative ratio when there are two corpuses. Or you can use a methodology such as Auto ML to let the computer automatically find the ratio.

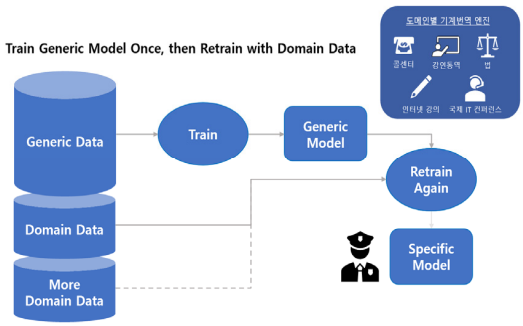

Fig.1. Methodology for Domain Specific Neural Machine Translation

3 Conclusion

This paper redefined the methodology of making a domain-specific machine translator. Domain-specific machine translation research can be extended to other domains and can be mainly used to improve the performance of deep learning-based natural language processing models. Through this research, it is possible to provide a machine translation model applying various styles and domains, so that it can be customized to the natural language processing service used in the domain, and it is based on deep learning technology, so it is a basic research of one of the related studies in the field. Can be used as.

4 Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2020-0-01819) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

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. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
5. Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. arXiv preprint arXiv:1905.02450.