A New Tool for Efficiently Generating Quality Estimation Datasets

Sugyeong Eo, Chanjun Park, Jaehyung Seo, Hyeonseok Moon, Heuiseok Lim[†]
Department of Computer Science and Engineering, Korea University
{djtnrud, bcj1210, seojae777, glee889, limhseok}@korea.ac.kr

Abstract

Building of data for quality estimation (QE) training is expensive and requires significant human labor. In this study, we focus on a data-centric approach while performing QE, and subsequently propose a fully automatic pseudo-QE dataset generation tool that generates QE datasets by receiving only monolingual or parallel corpus as the input. Consequently, the QE performance is enhanced either by data augmentation or by encouraging multiple language pairs to exploit the applicability of QE. Further, we intend to publicly release this user friendly QE dataset generation tool as we believe this tool provides a new, inexpensive method to the community for developing QE datasets.

1 Introduction

Quality estimation (QE) is the process of predicting the quality of machine translation results through source sentence and machine translation (MT) output [10]; recently, it has garnered a significant research interest [2, 5, 11]. Although a reference sentence is not required in QE, quality annotations according to the sentence or word level and human post-edited sentences are required to produce data for QE training [8]. In language selection for QE model construction, a large degree of dependency has been observed in translation experts that are proficient in both language pairs when undergoing correction.

In this paper, based on Eo et al. [3], we propose a fully automatic pseudo-QE dataset generation tool to address the limitation in the data construction aspect of QE. The tool is designed to be applicable to both monolingual corpus in the target language and parallel corpus configured with both source and target language. This significantly reduces the cost of building a QE dataset as it minimizes the human input and is easy to use through automated processes. In the case where there is a certain amount of constructed QE corpora, the tool presented in this paper can be used as a data augmentation technique for QE model training. Various applications of QE can be leveraged with pseudo-QE datasets created by the tool, even if there are no existing QE data in a particular language pair, especially in a low-resource setting [4, 7, 9].

2 Data Construction Process and Tool

This tool presents a total of three requirements from the user as needed. The first allows the user to select the language pair; the second to select the level (*i.e.,* word, sentence) at which the QE dataset is to be configured as an annotation; and the third to select either monolingual or parallel corpus. The QE dataset is produced through the process according to the user's choice. We visualize the overall steps for processing the pseudo-QE dataset generation in Figure 1.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

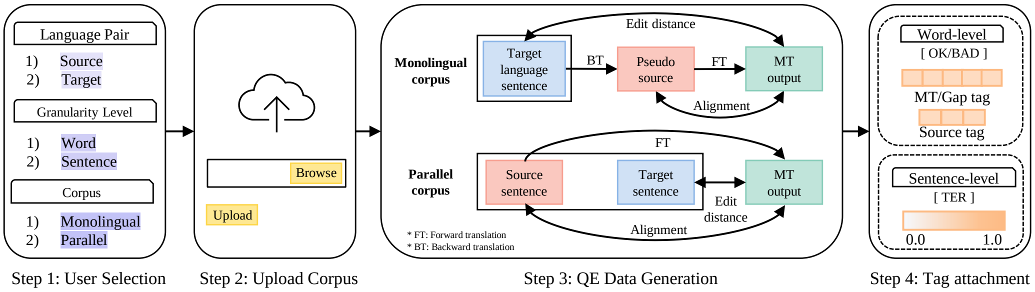

Figure 1: Overall process of building word- and sentence-level QE corpus based on our proposed tool.

Process

- STARTING FROM MONOLINGUAL CORPUS: The method of creating a QE dataset using a monolingual corpus is based on round trip translation and comprises four processes. Here, the monolingual corpus comprising the user's target language input serves as a post-edited sentence (*i.e.,* pseudo-reference sentence).

  For this corpus, the first process performs a backward translation to the source language to produce a pseudo-source sentence. In the second process, this pseudo-source sentence consists of an input to the forward translation process, and after the round trip, the MT output in the target language is obtained with errors attached. In the third process, the edit distance is measured to create an accurate sentence through minimal insertion, deletion, and substitution. Based on this, the final process labels the tag according to the granularity of QE level selected by the user. At the sentence-level, translation error rate (TER) [6], which is a ratio between the edit distance and pseudo-reference sentence, is scored. At the word-level, due to the need for source tag, an OK/BAD annotation is generated for each token after alignment is performed between the results of the backward translation and MT output.
- STARTING FROM PARALLEL CORPUS: When using the parallel corpus, a QE dataset is generated in three processes, with the target sentence in the parallel corpus acting as the pseudo-reference in the process of measuring the edit distance. In the first process, the source sentence is configured as an input during forward translation to the target language, and in the second process, the edit distance between this result and pseudo-reference sentence is measured. As a final process for tag attachment, edit distance was used to measure the TER score in the sentence-level. At word-level, the source, MT and gap tag is labeled based on edit distance after performing alignment with the source sentence and MT output.

Tools We configured the tools with accessible web applications that allow users to easily create QE datasets, which are publicly available [1]. The users have options for language pair, type of corpus, and granularity level, all of which allow them to build their own personalized QE dataset.

Our webserver is Flask-based, and we use a Google machine translator as a translation model as it is easy to use and has many current users. There are no language restrictions within the language pairs supported by Google translation. In the process of producing the QE data, tercom[6] was used to calculate the TER score, and in the case of alignment, the tool provided by Dyer et al. [1] was used. Open tool[2] released by Unbabel was used to generate word-level tags based on the edit distance.

3 Conclusion

In this study, we proposed an automated tool for generating pseudo QE datasets in an easy and inexpensive manner. This tool can increase the productivity in QE dataset generation and reduce the language pair constraint. In the future, we plan to combine data filtering to enhance the quality of the pseudo-QE dataset.

[1] http://nlplab.iptime.org:9091/

[2] https://github.com/Unbabel/word-level-qe-corpus-builder

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and IITP grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).Thanks to Seungjun Lee for helping us build the Tool. Heuiseok Lim† is a corresponding author.

References

[1] Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

[2] Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2021. Comparative analysis of current approaches to quality estimation for neural machine translation. *Applied Sciences*, 11(14):6584.

[3] Sugyeong Eo, Chanjun Park, Jaehyung Seo, Hyeonseok Moon, and Heuiseok Lim. 2021. Dealing with the paradox of quality estimation. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–10.

[4] Dongjun Lee. 2020. Cross-lingual transformers for neural automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 772–776.

[5] Joao Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André FT Martins. 2020. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036.

[6] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

[7] Lucia Specia. 2011. Exploiting objective annotations for minimising translation post-editing effort. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

[8] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

[9] Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

[10] Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

[11] Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, Yimeng Chen, et al. 2020. Hw-tsc's participation at wmt 2020 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1056–1061.