A Quality Estimation and Quality Evaluation Tool for the Translation Industry

Elena Murgolo
Orbital 14
Milan, Italy
emurgolo@orbital14.ai

Javad Pourmostafa
TSHD, CSAI Department,
Tilburg University
Tilburg, The Netherlands
j.pourmostafa@uvt.nl

Dimitar Shterionov
TSHD, CSAI Department,
Tilburg University,
Tilburg, The Netherlands
d.shterionov@uvt.nl

1 Introduction

With the increase in machine translation (MT) quality over the latest years, it has now become a common practice to integrate MT in the workflow of language service providers (LSPs) and other actors in the translation industry. With MT having a direct impact on the translation workflow, it is important not only to use high-quality MT systems, but also to understand the quality dimension so that the humans involved in the translation workflow can make informed decisions. The evaluation and monitoring of MT output quality has become one of the essential aspects of language technology management in LSPs' workflows. First, a general practice is to carry out human tests to evaluate MT output quality *before* deployment. Second, a quality estimate of the translated text, thus after deployment, can inform post-editors or even represent post-editing effort. In the former case, based on the quality assessment of a candidate engine, an informed decision can be made whether the engine would be deployed for production or not. In the latter, a quality estimate of the translation output can guide the human post-editor or even make rough approximations of the post-editing effort. Quality of an MT engine can be assessed on document or on sentence level. A tool to jointly provide all these functionalities does not exist yet.

While human evaluation is considered the most reliable method of analyzing MT quality, it is time-consuming, expensive, and hardly scalable. Human testing is also difficult to apply for actual projects during a production workflow. While some commercial products that can partly replace human testing already exist, they are usually CAT-dependent and cannot be employed independently from other language technology tools.

The overall objective of the project presented in this paper is to develop a machine translation quality assessment (MTQA) tool that simplifies the quality assessment of MT engines, combining quality evaluation and quality estimation on document and sentence level. To address both use cases, i.e., before general deployment and to estimate each translation's quality, this tool will comprise two working modes: a machine translation quality evaluation (MTQEv) and a quality estimation (MTQE) modes.

This 6-month project is a collaboration between Tilburg University and Orbital14, an R&D company owned and 100% financed by Italian LSP Aglatech14, whose funding is making this tool's development possible.

2 MTQA Tool Overview

The MTQA is designed as a standalone tool and an API that can be used by users or invoked by other tools. Behind the user interface lies a distributed architecture which operates in two modes. Intermediate and final results are displayed to the user; final results are made available for download.

MTQEv is a human-driven quality assessment module, in which one or more MT systems' quality is evaluated based on a human-generated translation reference. MTQEv is typically used to compare already-in-use and new MT models by means of comparing their translation to the human gold standard, using automatic metrics, such as TER (Snover et al., 2006), BLEU (Papineni et al., 2002), chrF (Popović, 2015) and others. In our MTQA tool, this mode shall be used to take an informed decision on a business level about the models to be deployed in production for different

language combinations and domains.

MTQE (machine translation quality estimation) is the process of predicting the quality of an MT system without human intervention or reference translations. MTQE can be at a word, sentence, or document level. In the case of document and sentence level, which are of interest for our project, the task is typically to predict a score that corresponds to a target evaluation criteria or metric. MTQE is the second mode the tool will be able to work in. Instead of comparing the MT output with an existing human translation, this mode will be used at the beginning of each translation project to evaluate the quality of the output by predicting the approximate number of changes a given MT output should undergo to reach acceptable quality. This mode shall be used to evaluate the usability of MT models for each project, in order to choose the best possible starting point for the PE (post-editing) and therefore to better allocate time and resources.

Both modes will be able to work both on document level and on segment level, so that the end-users, e.g., the project managers, will be able to choose the level of granularity they want to get to take a well-informed decision. To facilitate the use of the tool across all business workflows and for each use case, both modes will be integrated and independent from the other language technology tools that LSPs usually work with.

For MTQEv mode we employ the metrics TER, BLEU and chrF. For MTQE we first build neural QE models with the data described in Section 3; we then employ these models to score input data.

3 Working with industry data

An LSP could use either publicly available MT engines, or proprietary MT engines, trained specifically for the given translation use case, which would employ data, usually provided by the LSP to train a domain-specific and use-case-specific MT engine with highest quality.

The data an LSP usually translates is proprietary and cannot be publicly accessible, even for research purposes. A collaboration such as the one this project is based on, between Orbital14 and Tilburg, allows researchers and industry to work together on real use cases and proprietary data. Within the scope of this project, we exploit data that has been translated via trained and generic MT engines and was post-edited by Aglatech14 in the context of several translation projects. These data allow us to experiment with and build effective QE models that can be employed in the MTQA tool.

For this project we employ English–Italian data from the patent domain.

To this end, two types of data were provided by Aglatech14: (i) data that had been post-edited, in which case three documents were provided, source, MT output, and a post-edited version of the output (s-mt-pe); and (ii) data that had been translated by professional human linguists in the original project, in which case source and (human) translation (s-t) were provided.

To train our QE models we employed three different data sets: (i) the s-mt-pe documents; (ii) the s-t documents for which we translated the source using Aglatech14's MT engines and generated an s-mt-pe* corpus; and (iii) the open data sets BinQE (Turchi and Negri, 2014) and eSCAPE (Negri et al., 2018). For all data we computed the TER score between the MT and the post-edited or translated reference. This we used as target labels for our MTQE models.

References

Negri, Matteo, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, page 311–318. Association for Computational Linguistics, July 6-12.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, September 17-18.

Snover, Matthew, Bonnie Dorr, Rich 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, August 8–12.

Turchi, Marco and Matteo Negri. 2014. Automatic Annotation of Machine Translation Datasets with Binary Quality Judgements. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, May 26-31.