QUARTZ: Quality-Aware Machine Translation

José G. C. de Souza[1] **Ricardo Rei**[1,2,4] **Ana C Farinha**[1]
Helena Moniz[1,4,5] **André F. T. Martins**[1,2,3]
[1]Unbabel [2]Instituto Superior Técnico [3]Instituto de Telecomunicações
[4]INESC-ID [5]Faculdade de Letras da Universidade de Lisboa
Lisbon, Portugal
jose.souza, ricardo.rei, catarina.farinha, helena.moniz, andre.martins@unbabel.com

Abstract

This paper presents QUARTZ, QUality-AwaRe machine Translation, a project led by Unbabel and funded by the ELISE Open Call[1] which aims at developing machine translation systems that are more robust and produce fewer critical errors. With QUARTZ we want to enable machine translation for user-generated conversational content types that do not tolerate critical errors in automatic translations. The project runs from January to July 2022.

1 Introduction

Despite the progress in the fluency of machine translation (MT) systems, critical translation errors are still frequent, including deviations in meaning through toxic or offensive content, hallucinations, mistranslation of entities with health, safety, or financial implications, or deviation in sentiment polarity or negation. These errors occur more often when the source sentence is out of domain or contains typos, abbreviations, or capitalized text, all common with user-generated content. This lack of robustness prevents the use of MT systems in practical applications where the above errors cannot be tolerated.

QUARTZ aims to build reliable, quality-aware MT systems for user-generated conversational data. The project will address the limitations above by: (a) developing quality metrics capable of detecting critical errors and hallucinations; (b) endowing MT systems with a confidence (quality) score, and fine-tuning pre-trained MT models to the domains in which they will be used through quality-driven objectives.

This will be done by leveraging post-edited data and quality annotations produced by the Unbabel community and building upon the state-of-the-art, open-source quality estimation technology already existing at Unbabel: OPENKIWI (Kepler et al., 2019) and COMET (Rei et al., 2020). From a product perspective, focus will be given to conversational, user-generated data in a multilingual customer service scenario (email or chat involving a customer and an agent), in which Unbabel has renowned expertise and existing technology validated by existing customers. The solution aims to eliminate language barriers in the highly multilingual European market.

2 MT and Translation Quality

The current state of the art in MT is based on autoregressive sequence-to-sequence models trained with maximum likelihood and teacher forcing. This objective encourages the model to assign high probability to reference translations, but does not account for the severity of translation mistakes of the hypotheses generated. This leads to exposure bias, vulnerability to adversarial attacks, and no control for hallucinations, harmful content, and biases (Wang and Sennrich, 2020), hampering the responsible use of NMT for user-generated conversational content.

Project Overview Qualitative evaluation carried out by translators (post-editors and annotators) provides a human feedback loop that can generate large amounts of data with information about translation errors, their severities, and detailed quality annotations. The main methodology used to evaluate translations according to different aspects of translation quality is the industry-

[1] Funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 951847

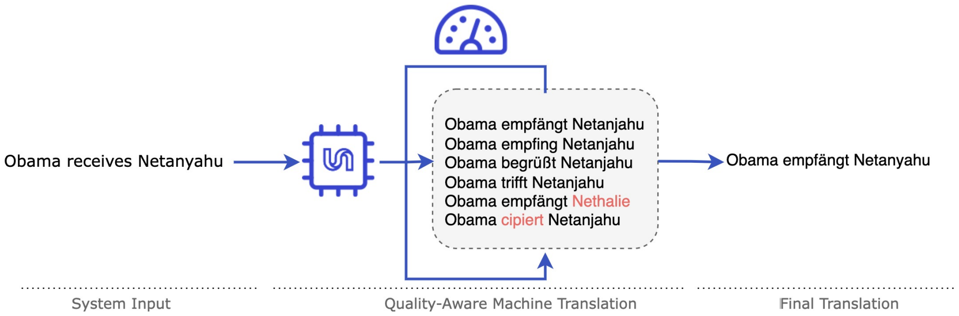

Figure 1: In QUARTZ quality estimation systems will interact directly with the machine translation system during the decoding phase to avoid critical errors. Words marked in red are considered errors.

adopted multi-dimensional quality (MQM) taxonomy (Lommel et al., 2014). Unbabel uses this data to train its open-source COMET and OPENKIWI frameworks to develop systems for MT evaluation and quality estimation, with MQM annotations and post-edits becoming a standard in Metrics and Quality Estimation WMT shared tasks (Freitag et al., 2021; Specia et al., 2021).

This project will close this loop by making MT systems quality-aware and robust. Decoding strategies for MT will be developed using the quality estimation metrics trained on the target domain data. The incorporation of these quality objectives into the decoding step of MT systems can have a big impact on controlling their tendency to produce hallucinations and other critical mistakes. This rationale is depicted in Figure 1.

Related Work Prior work on minimum Bayes risk (MBR) decoding paves the way to tune MT systems towards a given metric, but so far this has been done with purely lexical metrics such as BLEU (Müller and Sennrich, 2021) or neural metrics that do not capture severity and biases (Freitag et al., 2022). The main difference between QUARTZ and previous work is going beyond lexical metrics in incorporating quality scores for generating automatic translations.

References

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November. Association for Computational Linguistics.

Freitag, Markus, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. In *Accepted at Transactions of the Association for Computational Linguistics, presented at North American Chapter of the Association for Computational Linguistics 2022*, Seattle, Washington. Association for Computational Linguistics.

Kepler, Fabio, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy, July. Association for Computational Linguistics.

Lommel, Arle, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: Tecnologies de la Traducció*, 0:455–463, 12.

Müller, Mathias and Rico Sennrich. 2021. Understanding the properties of minimum bayes risk decoding in neural machine translation. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Specia, Lucia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online, November. Association for Computational Linguistics.

Wang, Chaojun and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online, July. Association for Computational Linguistics.