Robust, interpretable and efficient MT evaluation with fine-tuned metrics

Ricardo Rei
Instituto Superior Técnico, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
INESC-ID HLT lab, Rua Alves Redol, 9 1000-029 Lisboa, Portugal
Unbabel Research, R. Castilho 52, 1250-069 Lisboa, Portugal
Supervisors: Luísa Coheur and Alon Lavie
ricardo.rei@tecnico.ulisboa.pt

With the increasing need for Machine Translation (MT) in a world which is becoming globalized, there is also an increasing need to constantly evaluate the quality of the produced translations. This evaluation can be achieved through human annotators performing quality assessments or by employing automatic metrics. While human evaluation is preferred, it is expensive and time-consuming. Consequently, over the past decade, MT progress has primarily been measured using automatic metrics that assess lexical similarity against reference translations. However, numerous studies have demonstrated that lexical-based metrics do not correlate well with human judgments, casting doubt on the reliability of research in MT. Motivated by these challenges, the main goal of this thesis was to improve the current state of MT evaluation by developing new automatic metrics that satisfy four criteria: 1) strong correlation with human judgments, 2) robustness across different domains and language pairs, 3) interpretability, and 4) efficiency.

Based on recent advancements in cross-lingual language modeling, we hypothesize that a supervised metric incorporating the source-language input into the evaluation process will produce a more accurate MT evaluation. To validate this hypothesis, we introduce COMET (Crosslingual Optimized Metric for Evaluation of Translation), a neural framework for training multilingual MT evaluation models that serve as metrics. Models developed within the COMET framework are trained to predict human judgments of MT quality, such as *Direct Assessments* (DA), *Multidimensional Quality Metrics* (MQM), or *Human-mediated Translation Edit Rate* (HTER). Our results demonstrate that metrics developed within our framework achieve state-of-the-art correlations with human judgments across various domains and language pairs.

Nevertheless, lexical metrics still possess redeeming qualities in terms of interpretability and lightweight nature. In contrast, fine-tuned neural metrics like COMET are considered “slow black-boxes”. To address this, we employ neural explainability methods to reveal that these metrics leverage token-level information directly associated with translation errors. We showcase their effectiveness for interpreting state-of-the-art fine-tuned neural metrics by comparing token-level neural saliency maps with MQM annotations. Additionally, we present several experiments aimed at reducing the computational cost and model size of COMET while maintaining its state-of-the-art correlation with human judgments, thus bridging the performance gap between lexical and model-based metrics. That work, titled COMETINHO: THE LITTLE METRIC THAT COULD, was recognized with the Best Paper Award at EAMT 2022.

Realizing that system-level MT metrics alone are insufficient for comprehensive evaluation, this thesis also presents MT-TELESCOPE, a contrastive analysis tool that provides fine-grained segment-level insights into MT quality. By identifying the factors behind system performance, MT-TELESCOPE enables a deeper understanding of translation accuracy at the phenomenon level (e.g., named entities).

Over the past years, my thesis work has significantly influenced the field, inspiring research on quality-aware decoding – a paradigm that closely aligns with recent advances in test-time compute for large language models. By introducing high-performing, interpretable, and efficient evaluation metrics, my thesis work represents a substantial step forward in MT evaluation and has set a new standard for assessing translation quality. Receiving the EAMT 2022 Best Paper Award along with the Best Thesis Award at EAMT 2024 is a great honor and further solidifies the strength and recognition of my work in MT by the EAMT organizers.