Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Natural Language Processing (WATANABE TARO (Professor))		
Student ID	2311417	Submission date	2025 / 7 / 22
Name	LIU JINGXUAN		
Thesis title	Cross-lingual Scale: An Ignored Role in Multilingual Translation Evaluation		

Abstract

With the rapid development of Multilingual Neural Machine Translation (MNMT) systems, translation quality has been significantly improved, making it a key challenge to fairly and accurately evaluate the performance of MNMT systems, including Large Language Models (LLMs). Existing automatic evaluation methods usually represent the multilingual translation performance by directly averaging the scores of all language pairs, but this strategy has limitations. Because we believe that automatic evaluation metrics may show a " preference " for a specific language pair, resulting in significant differences in scoring ranges across languages and unfair cross-lingual evaluation. To verify this issue, we utilize GPT-4o to simulate the Multidimensional Quality Metrics (MQM) mechanism, construct a parallel multilingual dataset covering nine translation directions to compensate for the inadequacy of the existing dataset in terms of language coverage, and conduct a systematic investigation into the cross-lingual evaluation fairness of the mainstream automatic evaluation metrics. The results show that existing metrics generally suffer from inconsistent cross-lingual scoring scales. To this end, we propose three normalization-based framework to map the raw scores onto a uniform normalized scale, including 1) Language-specific quality level-wise normalization; 2) Language-specific global normalization; and 3) Global normalization across languages.The experimental results show that Language-specific global normalization can alleviate the inconsistency of cross-lingual scoring scales to a certain extent, which provides an idea to mitigate the problem.