Reliable Evaluation Metrics for Grammatical Error Correction

Name: Takumi Goto

Laboratory's name: Natural Language Processing Laboratory

Supervisor's name: Taro Watanabe

Abstract

Grammatical Error Correction (GEC) systems have evolved as writing support tools primarily targeting language learners. While the emergence of Large Language Models is rapidly expanding the range of available systems, identifying the best system remains a critical practical requirement. To this end, automatic evaluation metrics are essential. However, existing metrics face challenges across multiple dimensions, undermining the reliability of evaluation results. This dissertation aims to resolve these issues by addressing them through four perspectives: implementation, evaluation processes, vulnerability, and explainability. First, we identify that inconsistent interfaces in the implementations of existing metrics hinder fair meta-evaluation. To address this, we propose GEC-METRICS, a library designed with a unified interface. Second, although automatic evaluation aims to approximate human evaluation, we highlight a discrepancy between their evaluation processes: human evaluation is relative, while automatic evaluation is absolute. We demonstrate that bridging this gap improves correlation with human judgments across various metrics. Third, we reveal the existence of vulnerabilities to adversarial inputs, particularly in reference-free metrics. These vulnerabilities obstruct the appropriate selection of GEC systems. We demonstrate the feasibility of attacking existing metrics and show that metric ensembling serves as an effective short-term mitigation strategy. Finally, the low explainability issue exists in neural-based metrics because they typically output a single scalar value. We propose an edit-level feature attribution method and explain metric's internal decision at the edit level.