Incorporating Human-like Criteria into Automated Evaluation

Name Kosuke Doi

Laboratory's name Natural Language Processing Laboratory

Supervisor's name Taro Watanabe

Abstract

Evaluating generated sentences or texts—whether produced by humans or machines—is highly important in many fields, including education and natural language processing. Such evaluations are typically performed by humans but require significant time and effort. Automated evaluation offers a way to address the problem, with various metrics and models designed to align with human evaluation. Those metrics and models have been developed to achieve a higher correlation with human evaluation, but it is unclear whether the evaluation aspects used by humans are considered in the calculation of automatic evaluation scores. In this dissertation, we present ways to incorporate human-like criteria into automated evaluation in two different tasks: (1) essays writing and (2) simultaneous interpreting.

First, in evaluating essays, human raters consider grammatical items and their difficulties used in essays, while it is unclear whether state-of-the-art automated essay scoring (AES) models, which use BERT-based essay representations, capture these aspects. We propose ways to incorporate grammatical features into BERT-based AES models. We further use Item Response Theory to consider characteristics of individual grammatical items including their difficulties. Secondly, in simultaneous interpreting, especially for language pairs whose word order is different, human interpreters produce monotonic translations, which follow the word order of the source language. However, current automated evaluation metrics and models rely on written translation data that typically contain long-distance word reordering. We analyze the characteristics of monotonic translations, and use them as well as existing test sets for evaluating output from speech translation and simultaneous speech translation models.

The experimental results in the above two tasks provide empirical evidence to support effectiveness of incorporating human-like criteria into automated evaluation.