論文内容の要旨

申請者氏名　　土肥　康輔

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-interpretable 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. This work proposes ways to incorporate grammatical features into BERT-based AES models. Item Response Theory is employed to take into account 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. This work analyzes the characteristics of monotonic translations, and uses 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-interpretable criteria into automated evaluation.

論文審査結果の要旨

申請者氏名　　土肥　康輔

Neural machine translation (NMT) has achieved sufficient translation quality in the general domain, but not yet in the out-of-domain. Therefore, post-editing (PE), which manually corrects mistranslations, is still crucial, especially in fields where errors are not allowed, e.g., the medical domain. This dissertation tackles these essential problems of machine translation in terms of text generation and post-editing using interpretable models.

The first work prevents the degradation of the translation quality in the out-of-domain. In prior work, kNN (k-nearest neighbor) was employed in NMT models and applied to various domains using the example-based approach; however, the example search is time-consuming and the decoding speed becomes two orders of magnitude slower than that of standard NMT. To improve the decoding speed of kNN-MT, subset kNN-MT reduces the search space to the neighboring examples of the input sentence and employs an efficient computation method using the distance lookup table. The second work aims to improve the efficiency of human PE. Prior automatic PE (APE) models attempt to correct the outputs of an MT model; however, many APE models are based on sequence generation, and thus their decisions are harder to interpret for human post-editors. The new edit-based PE model breaks the editing process into two steps, “error detection” and “error correction.” The detector model tags each MT output token whether it should be corrected and/or reordered while the corrector model generates corrected words for the spans identified as errors.

The two studies are published as a high quality peer-reviewed journal paper and a peer-reviewed international conference paper. The research would have an impact not only to natural language processing, in particular, language education or machine translation, but to the relevant fields of machine learning as a way to incorporate human interpretable judgement in automatic evaluation. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.