Abstract

With the wide spread of the Internet, nowadays, many people make reservation for accommodation at hotels via websites. Furthermore, on many online hotel reservation websites, users can post reviews about a hotel at which they stayed and hotels can reply to those user reviews. It is important for hotels to reply to user reviews in order to protect their brand image and enhance customer's satisfaction. However, replying to a large number of reviews requires vast time and resources, and replying to negative reviews can impose a substantial mental burden on a hotel staff. Therefore, it is required to develop a system that can automatically generate replies to hotel reviews.

Most previous studies on automatically generating replies to reviews focus on reviews about applications(software) and products, while generation of replies to hotel reviews has not been investigated much. In addition, previous studies do not pay attention to generation of replies that comprehensively address multiple aspects in reviews. Here "aspect" of a hotel refers to an element of a hotel such as "room", "bath", "staff" and other facilities, functions, and services, which is often evaluated in a review. If a user expresses dissatisfaction with multiple aspects but a hotel addresses only a few aspects or no aspect in a reply, the user may feel that the reply is insincere and get a negative impression on the hotel.

This study aims to generate appropriate replies to low-rated hotel reviews. Especially, the following two characteristics are heavily considered. First, when a user expresses dissatisfaction with multiple aspects of a hotel, the generated reply should address all these aspects. Second, generated replies should be diverse, not consist of only generic sentences in other words. Here, a generic sentence is defined as a sentence that only conveys general meaning and can be used in many reviews, such as "thank you." and "we are looking forward to your next visit." We aim to generate not generic sentences but specific sentences that are highly coherent with user's complaints.

The procedures of the proposed method are as follows. First, a hotel review is divided into individual sentences. Next, "complaint classification model" judges whether each sentence is a complaint, and sentences not identified as complaints are excluded. Then, the remaining complaint sentences are fed into "reply generation model" to generate a reply to each of the input sentences. Finally, the reply sentences generated for all complaint sentences are concatenated to produce the final reply to the review.

The complaint classification model is trained using Bidirectional Encoder Representations from Transformers(BERT). Reviews that consist of a single sentence and are labeled as a complaint are extracted from the Rakuten Travel dataset as positive samples, while the same number of single-sentence reviews without com-

plaint labels are randomly chosen as negative samples. These samples are used as the training data for fine-tuning of BERT.

The reply generation model is trained by fine-tuning the Bidirectional Auto-Regressive Transformer (BART), which is a sequence-to-sequence model. As the training data, pairs of user reviews and their corresponding replies written by hotels are obtained from the Rakuten Travel dataset. To improve the quality of the training data, two filtering methods are employed. The first filtering is to remove replies that do not address aspects in user reviews. Only pairs of a review and a reply where both contain the same aspect word are kept, while others are removed. This ensures that the reply generation model can produce replies that address the aspects appeared in the review. The aspect words of hotels are extracted from the user reviews in the training data in advance. TF-IDF scores of all words in all reviews are calculated, and the average of the top 1000 TF-IDF scores is used as the score of each word. The top 500 words ranked by this score are chosen as the aspect words.

The second filtering is to remove generic sentences from the training data. This process is expected to prevent the reply generation model from generating stereotyped replies. The detailed procedures of this filtering are as follows. Replies of hotels in the training data are divided into individual sentences, and a generic score is calculated for each sentence. This score is derived from the average of the frequency of all word tri-grams in the sentence, where the frequency is the number of occurrence of the word tri-gram in the training data. The top 30% sentences with the highest generic score are considered as generic and are removed from the training data.

In the last step of concatenation of replies to individual review sentences, duplicate replies are excluded. The similarity between the generated reply sentences is measured using a normalized edit distance. If this value is 0.1 or lower, the reply sentence that appears later in the original review is retained, and the other sentence is discarded. However, replies containing aspect words are always kept. After this procedure, the remaining reply sentences are concatenated in the same order as in the original review to form the final reply.

Several experiments were carried out to evaluate each of the complaint classification model and the reply generation model. To evaluate the complaint classification model, the model was applied for the classification of whether each review in the test data was a complaint, and the accuracy, precision, recall, and F1-score were measured. The best results were obtained when the number of training epochs was set to one, yielding 0.8877 accuracy, 0.8718 precision, 0.9091 recall, and 0.8901 F1-score. These results indicated that the performance of the complaint classification model was sufficiently high.

In the evaluation of the reply generation model, we compared several methods, such as methods with and without two filtering methods, and methods that generated replies to individual sentences and concatenated them or generated one reply to an entire review. A model fine-tuned from the training data without our filtering methods was used as a baseline. First, an automatic evaluation was conducted. The quality of the generated replies was assessed by BLEU-4 and DISTINCT-4. The BLEU-4 score of the baseline was higher than that of the proposed methods. The baseline tended to generate generic sentences, but the reference replies in the dataset also contained many generic sentences. Thus, many word n-grams were overlapped in the generated reply and reference, causing the BLUE-4 score to become high. Besides, the DISTINCT-4 score of the proposed method with filtering was higher than that of the baseline, indicating that our filtering methods were effective to suppress generation of generic sentences. Next, we conducted a human evaluation. Seven subjects rated the fluency, non-redundancy, overall quality, and aspect coverage rate (proportion of aspects appearing in reviews that are addressed in replies) of the generated replies. The results showed that the proposed method, which filtered out replies not addressing aspects, had a higher aspect coverage rate compared to the baseline, but its fluency and non-redundancy were worse. It indicates there is a trade-off between the aspect coverage rate and the fluency/non-redundancy. Filtering of generic sentences improved the non-redundancy, as well as the fluency and the overall quality. It means that the generation of stereotyped expressions is suppressed. In addition, for reviews containing multiple aspects, the technique to generate replies to not an entire review but individual sentences could improve both the aspect coverage rate and the overall quality, showing its effectiveness.