Generation of Diverse Responses to Reviews of Accommodations Considering Complaints about Multiple Aspects

Kiyoaki Shirai[1] **Yuta Murakoshi**[2] **Natthawut Kertkeidkachorn**[3]
[1,3]Japan Advanced Institute of Science and Technology
[2]KDDI Agile Development Center Corporation
[1]kshirai@jaist.ac.jp [2]yuta.murakoshi@gmail.com [3]natt@jaist.ac.jp

Abstract

It is important for a hotel manager to reply to customer reviews that complain about the services and facilities etc. of the hotel on an online booking website, in order to reduce the customer's dissatisfaction. However, it is rather hard to manually respond to all the aspects complained about in many reviews. This paper proposes a novel method to automatically generate a hotel's response to a given customer review, aiming to mention all the aspects complained about, in a wide variety of expressions. Two filtering methods of the training data are proposed: one is to remove responses that do not refer to an aspects in a review, the other is to remove general sentences with high frequencies in the training corpus. In addition, responses are separately generated for each of the sentences in a review, then they are integrated to form a final response. Our proposed method is assessed by automatic and human evaluation. The results show that both the filtering methods and the sentence-based generation can improve the quality of the generated responses.

1 Introduction

Nowadays, online reservation of accommodations has become popular, and websites for travelers are widely available. In many hotel booking websites, customers are able to not only compare hotels but also write a review after they stay at a hotel. In addition, the manager of the hotel can reply to a customer's review on the same website. Customers often express negative opinions and complaints about a hotel. It is important for a hotel to respond to such negative reviews in order to reduce customers' dissatisfaction and not to fall into disrepute. However, responding to many reviews imposes a heavy burden on a hotel manager. Therefore, the automatic generation of responses to customers' reviews, especially negative ones, is in great demand by hotels.

The goal of this paper is to automatically generate an appropriate response to a review including customer's complaints. Especially, the following two points are taken into account. One is consistency. A customer may express his/her complaints about two or more aspects of a hotel. Here, consistency means that the hotel refers to those aspects exhaustively in a response. For example, when a customer expresses complaints about the two aspects, "room cleaning" and "front desk," and a hotel manager does not apologize for one or both aspects, the customer will continue to feel dissatisfied with the hotel. Consequently, the hotel's response should mention all the aspects in the review. The other point is diversity. Neural text generation models tend to produce general expressions (Holtzman et al., 2020), generating short and stereotyped responses. A simple apology such as "We are sorry." or "We apologize to you for your trouble." is insufficient to satisfy a negative customer, since the customer feels such a naive response to be insincere. It is preferable to generate responses with various linguistic expressions. Therefore, our primary goal is to generate various (non-stereotyped) responses, which apologize for all aspects complained about in a given review.

Our proposed method is based on a common sequence-to-sequence (seq2seq) model that accepts a review as an input and generates a response as an output. To achieve our goal, we propose two filtering methods to improve the quality of the training data. We also propose an approach to split a review into sentences, generate responses for each of the sentences, and merge them so that explanations for all the aspects complained about are included in the response. The target language in this study is Japanese. The contributions of our paper are summarized as follows:

- We propose two methods to filter the training data so as to improve the diversity and consistency in the generation of a hotel's response.

- We propose a sentence-based generation approach to improve the consistency of the responses.
- We demonstrate the effectiveness of our proposed method by automatic and manual evaluation.

2 Related Work

Several studies have been made of the automatic generation of responses to a text in a website. Gao et al. (2019) propose RRGen, a system to automatically generate a response of a developer to a user review in an app store such Apple's App Store and Google Play. RRGen is based on an Encoder–Decoder model of a Recurrent Neural Network (RNN) where four features of a review (category of app, length of review, user's rating, and user's sentiment) are incorporated by an attention mechanism. By an ablation test, they demonstrate that each of four features can contribute to improving the quality of the generated responses. Zhao et al. (2019) generate a response of a customer service provider to a product review in an Electronic Commerce (EC) website. External information of a target product is incorporated into a seq2seq model by a gated multi-source attention mechanism and copy mechanism (Gu et al., 2016). Their model can generate sentences including information about the product, such as its brand and material, as real responses. Roy et al. (2022) aim to answer a user's question in a Question Answering (QA) platform in an EC website, and propose a method to retrieve relevant reviews for a given question, which may contain answers to the question.

Generation of responses in the hotel domain has also been studied. Kew and Volk (2022) focus on generating not a generic but a specific response that addresses the customer's comments in a hotel review. Three methods to remove generic responses from the dataset are proposed: (1) lexical frequency, which removes sentences including words with high frequencies, (2) sentence average, which discards sentences similar to prepared generic example sentences, and (3) language model perplexity, which filters out sentences with low perplexity calculated by a GPT-2 distilled for the hotel domain. After applying the above filtering methods to the training data, BART (Lewis et al., 2020) is fine-tuned as a response generation model. Using both automatic and human evaluation, they demonstrate that these three methods can contribute to the improvement of the specificity of the generated responses. Igusa and Toriumi (2021) generate responses to hotel reviews written in Japanese. An RNN seq2seq model is trained from a dataset of actual customer reviews and responses in the hotel booking website. In addition, to incorporate the information of the review into the model, embeddings of the rating by a reviewer and the length of the response are concatenated to the last hidden states of the encoder.

Kew et al. (2020) investigate what happens when moving to a different domain in response generation tasks. They extend Gao's model (Gao et al., 2019), developed for response generation in the app domain, and apply it to the hospitality domain (i.e., hotel and restaurant reviews). Results of their experiments show that the performance on the hospitality domain is much worse than that on the app domain. They determine that the major causes are the lengths of the reviews (reviews in the hospitality domain are much longer) and the textual variation in the responses (responses in the app domain are less diverse, thus easy to generate), and conclude that response generation in the hospitality domain is a more challenging task.

Unlike the previous studies on the generation of responses to hotel reviews, we mainly focus on generating an appropriate response to customers' complaints. An important characteristic of our method is to produce apologies for multiple aspects complained about in a review, with non-stereotyped expressions.

3 Proposed Method

Figure 1 shows an overview of our proposed method, where the input is a customer review and the output is the hotel's response to it. First, the review is split into sentences (§3.1). Second, each sentence is classified as to whether it contains a complaint, and sentences not including complaints are discarded (§3.2). Third, for each remaining sentence, a response is generated by a seq2seq model (§3.3). Finally, the generated responses are merged to form a final response (§3.4).

A straightforward approach to generating responses is to train an end-to-end model that accepts an original review and generates a response to it. However, such a model may often fail to mention all the complaints in the review, especially

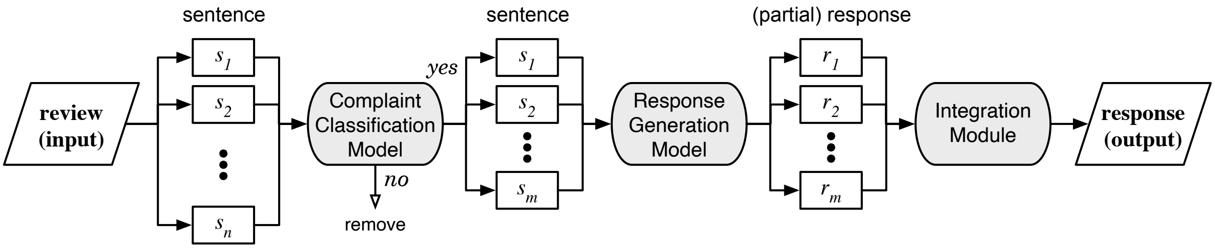

Figure 1: Overview of proposed method.

when the review is long and contains many complaints. Supposing that the complaints about multiple aspects appear in different sentences in the review, our method tackles this problem by generating responses from the individual sentences. This sentence-based generation approach enables us to reply comprehensively to complaints about multiple aspects.

Dataset Rakuten Data (Rakuten Institute of Technology) is used to train the response generation model and the complaint classification model. A part of Rakuten Data is a collection of customer reviews and responses to them by hotels, which are posted on the hotel booking website "Rakuten Travel." In addition, the reviews are annotated with a label that expresses a content of it, such as "complaint" and "impression". Hereafter, this dataset is called "Rakuten Travel dataset".

3.1 Sentence Split

The customer review is split into sentences by symbols indicating the end of a sentence such as a period ("."), question mark ("?") and exclamation mark ("!"). The obtained sequence of sentences is denoted by $S = (s_1, \cdots, s_n)$.[1]

3.2 Classification of Complaints

Since our main purpose is to reply to customers' complaints, sentences not containing complaints are removed. We train a binary classifier to judge whether a sentence expresses a customer's complaint. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is chosen as the complaint classification model. The BERT base Japanese (Tohoku NLP Group), which is trained from 17M sentences in Japanese Wikipedia, is used as the pre-trained model. It is fine-tuned using a labeled dataset. The sentences $s_i \in S$ are classified by the fine-tuned BERT model, then only the sentences classified as "yes" are added to a sequence of sentences $S_c = (s_1, \cdots, s_m)$, where $m \leq n$.

When the BERT model is fine-tuned, reviews labeled with the "complaint" tag in the Rakuten Travel dataset are used as the positive samples, and other reviews are used as the negative samples. In general, the reviews in the dataset are documents consisting of several sentences, while the complaint classification model is supposed to classify a single sentence. Therefore, only reviews containing one sentence are used. We make a balanced training dataset consisting of the same number of positive and negative samples. Since the number of the reviews labeled with "complaint" is smaller, first, all complaints are extracted, then an equal number of non-complaints are randomly chosen.

3.3 Generation of a Response

A response is generated for each of the complaining sentences in S_c. Our response generation model is a seq2seq model that converts a single sentence in a review into a response to it. We use BART (Lewis et al., 2020) as the base model. Japanese BART base (Language Media Processing Lab at Kyoto University) is a pre-trained BART model for Japanese. It is fine-tuned using pairs of a review with the "complaint" tag and a response to it in the Rakuten Travel dataset. We denote a sequence of the generated responses by $R = (r_1, \cdots, r_m)$, where r_i is generated from s_i in S_c.

In the Rakuten Travel dataset, it is found that a considerable number of responses by hotel managers do not refer to anything about the customer's complaints. We eliminate such inappropriate samples to generate desirable responses as discussed in Section 1. More specifically, two filtering methods, aspect filtering and generality filtering, are applied to the training data before the fine-tuning of the

[1] In the Rakuten data, very few writers omit a punctuation mark at the end of a sentence. In these cases, the texts are treated as a single sentence. The proportions of such reviews and replies are 2.69% and 0.270% respectively.

BART model.

3.3.1 Aspect Filtering

One of our goals is to generate a response that mentions all the complaints about multiple aspects. The first filtering removes from the training data responses that do not mention any aspects about which a customer made a complaint.

First, A, a set of aspect terms in the hotel domain, is constructed from V, a set of reviews labeled with the "complaint" tag. In this study, domain specific keywords are extracted from V as the aspects. For each word w_i in V, a score of its salience in the hotel domain is calculated using Equation (1).

$$SA(w_i) = \mathrm{ave}_{r_j \in \mathrm{TOP}_{1K}(w_i)} \text{TF-IDF}(w_i, r_j) \quad (1)$$

Here, r_j is a review in V, TF-IDF(w_i, r_j) is TF-IDF of w_i in r_j where V is the entire document set, and $\mathrm{TOP}_{1K}(w_i)$ is a set of the top 1000 reviews ranked by TF-IDF$(w_i, *)$. That is, $SA(w_i)$ is the average of the 1000 top-ranked TF-IDF scores. Then, we choose 500 words whose $SA(w_i)$ is the highest to form a set of the aspects A. We confirmed that most of the extracted aspects were appropriate. Several examples are shown in Table 1, where the original Japanese words are translated into English.

parking, room, drain, reservation, cigarette, shower, odor, bathroom, hospitality, breakfast, towel, cleaning, air conditioner

Table 1: Example of aspects (English translations).

After obtaining A, each pair of a review and a response is removed if (1) no aspect appears in the review or (2) the same aspect $a_i \in A$ does not appear in both the review and response. This filtering ensures that a response in the training data mentions an aspect in the corresponding review.

3.3.2 Generality Filtering

Kew and Volk (2022) suppose that, in the training of the text generation model, general expressions that frequently appear in the training data are harmful and degrade the ability of the model to generate specific expressions. Following their idea, we propose the second filtering method that removes general and common sentences from the dataset in order to generate expressions that are not stereotyped, but varied.

First, the responses in the Rakuten Travel dataset are split into sentences. Next, for each sentence s_k, the score of its generality is calculated by

$$G(s_k) = \mathrm{ave}_{tg_i \in s_k} \mathrm{fre}(tg_i), \quad (2)$$

where tg_i is the ith word tri-gram in the sentence s_k, and $\mathrm{fre}(tg_i)$ is the frequency of tg_i in the training data. That is, the generality of s_k is considered to be high when it contains word tri-grams with high frequencies.

All the sentences are sorted in descending order of $G(s_k)$, and the top 30% of the sentences are removed. After the filtering, the samples in the training data are pairs of an original review and a response consisting of the remaining sentences. If all sentences in a response have been removed, those samples are removed from the dataset.

Table 2 shows examples of sentences that get removed, and their generality scores.[2] We found that most of the removed sentences were general and typical.

Sentence	Score
I'm terribly sorry.	78544
I sincerely apologize.	49978
Thank you very much for staying with us.	48500
We sincerely look forward to welcoming you again.	39078
We understand and accept your point.	34997

Table 2: Examples of removed general sentences (English translation).

3.4 Integration of Responses

After obtaining R, the m generated responses are merged into a single document as the final output. The responses are concatenated in the same order as the source sentences in the input review. Since the responses are independently generated, some sentences might be duplicated and redundant. Therefore, redundant sentences are removed before merging the responses. Specifically, if the normalized edit distance (Levenshtein, 1966) of two sentences is smaller than the pre-defined threshold (0.1 in this study), the first sentence is kept and the second sentence is removed in the order of the appearance of the source sentences in the review. However, sentences including the aspects are always kept.

[2] The original Japanese sentences are shown in Appendix A.

Algorithm 1 shows the pseudocode of the integration of the responses. Since each response r_i consists of two or more sentences in general, all r_i are split into sentences to make a list of the sentences S_R (line 1). The sentence s_i is added to the end of S_O (a list of the output sentences) if its minimum edit distance to the sentences that have already been selected as the output is greater than 0.1 or if it contains an aspect term, otherwise removed (lines 4–9). Finally, the final response (output) O is obtained by concatenating the sentences in S_O (line 11).

Algorithm 1 Pseudocode of integration module.

```
Input: $R = (r_1, \cdots, r_m)$    ▷ in the order of the
     source sentences in the input review.
Output: $O$
 1: $S_R \leftarrow \bigcup_{r_i \in R}$ split-to-sentences($r_i$)
 2: $S_O \leftarrow ( )$                        ▷ empty list
 3: for $i = 1$ to $|S_R|$ do
 4:     $d = \min_{s_j \in S_O}$ edit-distance($s_i, s_j$)
 5:     if $d > 0.1$ or $s_i$ contains aspect then
 6:         append($S_O$,$s_i$)       ▷ $s_i$ is added to $S_O$
 7:     else
 8:         ;                         ▷ $s_i$ is removed
 9:     end if
10: end for
11: $O \leftarrow$ concatenate($S_O$)
```

4 Evaluation

4.1 Dataset

The Rakuten Travel dataset is used for the experiments to evaluate our proposed method. To train and evaluate the complaint classification model, as described in subsection 3.2, the reviews consisting of a single sentence labeled with the "complaint" tag are used as the positive samples. Those tagged with other tags are used as the negative samples. The reviews are split into 80% training data and 20% test data. The statistics of the dataset are shown in Table 3.

	Positive	Negative	Total
Training	16,099	16,099	32,198
Test	4,025	4,025	8,050

Table 3: Dataset for complaint classification.

To train the response generation model, pairs of a review labeled with the "complaint" tag and the hotel's response to that review are extracted from the Rakuten Travel dataset. Although our response generation model is supposed to accept a single sentence as an input, the reviews consisting of not only one sentence but also multiple sentences are used. This is because more training samples are required to train the seq2seq model. The samples of the response generation are split into 90% training data, 5% development data, and 5% test data. The development data was used to investigate the filtering methods at the initial stage of this study. Table 4 shows the statistics of the dataset.[3] The two filtering methods decrease the number of the samples in the training data by 29%.

Data	# samples
Training	147,749
Training (after filtering)	105,241
Development	8,209
Test	8,209

Table 4: Dataset for response generation.

4.2 Evaluation of Complaint Classification Model

The model of the complaint classification is evaluated first. The BERT model is fine-tuned using AdamW (Loshchilov and Hutter, 2019). The number of the epochs is set to 1, the learning rate is set to $2\mathrm{e}^{-5}$, and the other hyperparameters are set to the default parameters of AdamW.[4]

The accuracy as well as the precision, recall, and $F1$-score of the "complaint" class are shown in Table 5. The accuracy and $F1$-score are 0.8877 and 0.8901, respectively, indicating that the performance of the complaint classification model is sufficiently high.

Accuracy	Precision	Recall	$F1$-score
0.8877	0.8718	0.9091	0.8901

Table 5: Results of complaint classification.

4.3 Evaluation of Proposed Method

4.3.1 Experimental Setting

In these experiments, the six methods in Table 6 and GOLD (the ground-truth response in the dataset) are compared. "BASELINE" is a method

[3] Additional statistics are shown in Appendix B.

[4] We also set the learning rate to $5\mathrm{e}^{-6}$ and $1\mathrm{e}^{-6}$ and found that all the trained models were comparable.

Method	Filtering		Sentence
	Aspect	Generality	Split
BASELINE	×	×	×
BASELINE-S	×	×	✓
PRO-A-S	✓	×	✓
PRO-G-S	×	✓	✓
PRO-AG	✓	✓	×
PRO-AG-S	✓	✓	✓

Table 6: Summary of response generation methods.

Method	BLEU-4	DISTINCT-4
BASELINE	**0.1233**	0.0313
BASELINE-S	0.1034	0.0224
PRO-A-S	0.0962	0.0562
PRO-G-S	0.0740	0.0395
PRO-AG	0.0660	**0.0585**
PRO-AG-S	0.0667	0.0533

Table 7: Automatic evaluation of response generation methods.

that simply uses the BART model for response generation. "PRO" indicates the variations of our proposed method. A response is produced by sentence-based generation in the methods with "-S", while a review is not split into sentences but the original review is fed into the model in the methods without "-S". The symbols "A" and "G" indicate that the aspect filtering (§3.3.1) and the generality filtering (§3.3.2) are applied, respectively.

When the pre-trained BART model is fine-tuned to obtain the response generation model, the hyper-parameters are set as follows: the number of the epochs is set to 5, the learning rate to 3e^{-5}, and the dropout rate to 0.3.

4.3.2 Automatic Evaluation

First, our methods and baselines are automatically evaluated. Two evaluation criteria are used: BLEU (Papineni et al., 2002) and DISTINCT (Li et al., 2016). BLEU evaluates how the generated response is close to the ground-truth, while DISTINCT evaluates the variety of the generated response. Specifically, BLEU-4 and DISTINCT-4 based on the word 4-grams are measured.

Table 7 shows the results of the automatic evaluation. Comparing the methods with and without the filtering, it is found that DISTINCT-4 is much improved by removing inappropriate samples from the training data. PRO-G-S outperforms BASELINE and BASELINE-S, indicating the effectiveness of the generality filtering to produce more diverse responses. We guess that the aspect filtering can also contribute to improve the variety, because most of stereotyped responses do not contain an aspect and can be removed by this filtering. This is supported by the fact that DISTINCT-4 of PRO-A-S is better than that of the baseline. In addition, the use of two filtering methods further improves the variety of the generated responses as the highest DISTINCT-4 is achieved by PRO-AG.

Besides, BLEU-4 of the methods with the filtering are worse than those without the filtering. The baseline methods often generate stereotyped sentences, and many actual responses by hotels in the dataset are also short, fixed and stereotyped. Thus many overlaps of the word 4-grams between the generated and gold responses are found, resulting in the high BLUE-4.

4.3.3 Human Evaluation

Human evaluation is also conducted. First, 50 reviews in the test data are randomly chosen. The quality of the responses to those reviews generated by the five methods is manually assessed.[5] GOLD is also evaluated for the comparison. Seven subjects, who are graduate students, are invited to the human evaluation. They are asked to evaluate the generated responses from the following points of view. The details of the instructions to the human subjects are shown in Appendix C. Note that each subject evaluates the responses to all 50 reviews.

Fluency To rate how natural a response is as a Japanese text.

Non-redundancy To rate how redundant a response is. A response where the same or almost similar expressions are repeated should be rated lower.

Overall Score To rate the overall score of a response. We instruct the subjects to answer this by: "Supposing you had written the complaints in the review, how would you feel about the hotel's response?"

Mention of aspect To check whether a response mentions aspects in a review. All aspects in a review are manually extracted before the assessment, and subjects are asked to answer "yes" or "no" for each aspect.

[5] BASELINE is omitted in the human evaluation to lighten the burden imposed on the human subjects.

(a) all reviews

Method	F	N-R	O	CoA
BASELINE-S	4.58	4.44	2.53	0.338
PRO-A-S	4.34^-	4.16^-	2.72*	0.581*
PRO-G-S	4.63	4.50	2.80*	0.415*
PRO-AG	4.66	4.71*	2.98*	0.450*
PRO-AG-S	4.48	4.17^-	2.77*	0.538*
GOLD	4.63	4.84	3.99	0.652

(b) only reviews containing multiple aspects

Method	F	N-R	O	CoA
BASELINE-S	4.54	4.27	2.48	0.296
PRO-A-S	4.27^-	3.99^-	2.58	0.473*
PRO-G-S	4.54	4.30	2.81*	0.329*
PRO-AG	4.60	4.73*	2.61	0.232
PRO-AG-S	4.50	3.97^-	2.73*	0.466*
GOLD	4.56	4.82	3.94	0.596

Table 8: Result of human evaluation. F, N-R, O and CoA stand for fluency, non-redundancy, overall score and coverage of aspect. The mark * or $^-$ indicates the method is significantly better or worse than BASELINE-S (by t-test, $p < 0.01$).

The fluency, non-redundancy, and overall score are rated on a five-point scale from 1 to 5. As for the mention of the aspect, we calculate "Coverage of Aspect" ("CoA" in short) defined by Equation (3) based on the subjects' answers.

$$\textbf{CoA} = \frac{\#\text{ of aspects mentioned in responses}}{\#\text{ of aspects in all reviews}} \tag{3}$$

Table 8 (a) shows the average of the criteria of the seven subjects. Fleiss' κ of the subjects is 0.34 for fluency, 0.62 for non-redundancy, 0.42 for overall score, and 0.77 for the number of mentioned aspects, indicating moderate agreement.

Aspect filtering The proposed method using the aspect filtering (PRO-A-S, PRO-AG-S) outperforms BASELINE-S in terms of the CoA, thus our aspect filtering can contribute to replying to all the aspects complained about. On the other hand, the values of F and N-R are decreased by using this filtering. This may be because the similar sentences are repeated by mentioning multiple aspects. Although redundant sentences are removed in our integration module (§3.4), similar sentences still remain. We can find a trade-off between the aspect coverage and the fluency/non-redundancy.

Generality filtering It is confirmed that the non-redundancy of the methods using the generality filtering is better than BASELINE-S. In addition, the fluency and overall score are also better. Therefore, the generality filtering can suppress the generation of stereotyped sentences and improve the quality of the generated responses. An exceptional case is that the non-redundancy of PRO-AG-S is worse than BASELINE-S. This may be due to the trade-off between N-R and CoA; the use of the aspect filtering in PRO-AG-S causes an increase of CoA but a decrease of N-R.

Sentence-based generation Comparing the methods with and without the sentence-based generation, the aspect coverage of PRO-AG-S is significantly better than that of PRO-AG. Several aspects may appear in different sentences in a review, thus generating responses from each of the sentences can include a thorough mention of each of those aspects. Besides, PRO-AG-S achieves worse fluency and overall score. Handling a whole review can generate a more fluent and less redundant response, while our sentence-based generation sometimes fails to generate natural sentences and avoid repetition. Since the overall score of PRO-AG-S is worse than PRO-AG, the contribution of the sentence-based generation is uncertain.

The sentence-based generation method is designed to mention all the aspects in a review. Note that not all reviews contain multiple aspects. Among 50 reviews in the test data, 22 reviews include two or more aspects. To clarify the effectiveness of the sentence-based generation, we measure the average scores of those 22 reviews. The results are shown in Table 8 (b). It is found that the CoA is more improved by the sentence-based generation (PRO-AG-S 0.466 vs. PRO-AG 0.232), and the overall score of PRO-AG-S is higher than PRO-AG. Therefore, the sentence-based generation is not adequate for a review including one aspect, but remarkably effective for a review including multiple aspects. This leads to a simple method to integrate the two approaches: (1) the number of aspects in a review is counted, (2) if there is only one aspect, PRO-AG is applied, otherwise PRO-AG-S is applied.

4.4 Case Study

Table 9 shows an original review and responses generated by BASELINE-S and PRO-AG-S, and

Review	Response		
	BASELINE-S	PRO-AG-S	GOLD
First of all, if all rooms are no-smoking, please make this clear. It might be mentioned somewhere on the site, but it's hard to find when skimming through. The staff had a pleasant demeanor, which was satisfying. The room design was good, but there was no bathtub. The private bath (the one at the far left) couldn't be temperature controlled, so it was like a punishment to have a hot bath in the middle of summer. The shower was too weak. Eventually, we had to scoop the hot water from the bathtub to use. The public bath was ordinary but good. The food aimed to bring out the natural flavors of the ingredients. To put it negatively, it didn't seem like much effort was put into it. However, there was one dish, the lemon-simmered radish (?), that was exceptionally delicious. This dish alone was impressive.	We apologize for the inconvenience caused by the temperature of the private bath. Our staff will make every effort to ensure our guests have a comfortable stay. Thank you very much for choosing our hotel.	We sincerely apologize for the inconvenience caused regarding the information on no-smoking rooms. We deeply regret any trouble caused by our website. We also apologize for the inconvenience related to the temperature adjustment of the private bath. We are very sorry about the issue with the shower. We will strive to ensure that such issues do not occur in the future.	This is XXX. Thank you very much for staying with us the other day. We will work on improving the areas you pointed out, starting with what we can address immediately. We appreciate your continued patronage of XXX. (XXX is the name of the hotel.)

Table 9: Examples of generated responses (English translation).

GOLD as examples of the response generation.[6] The reviewer complains about three aspects, "no-smoking" (it is not announced in the hotel website), "private bath," and "shower." On the one hand, in the response of BASELINE-S, not all the complaints of the reviewer are mentioned. The hotel apologizes only for the aspect "private bath." On the other hand, in PRO-AG-S, the hotel apologizes for the three aspects one by one, which might be more appropriate as a response. However, the response is somewhat redundant, since the apologies are repeated, as indicated by the wavy lines. Besides, the response of GOLD just expresses the stereotyped sentences.

5 Conclusion

This paper proposed a novel method to generate a hotel's response to a given review that expressed customer's complaints. The results of the experiments demonstrated that our proposed method was significantly better than the baseline in terms of the overall score and the coverage of the aspects.

Our method could appropriately reply to a review complaining about multiple aspects, but the response tended to be long and contain redundant sentences. In the future, we will explore ways to revise the response integration module to improve the non-redundancy and fluency. More sophisticated methods of measuring the similarity between sentences should be investigated to detect redundant sentences. Another important line of future research is to handle multiple aspects more appropriately. We suppose that one sentence contains one aspect, but two or more aspects can appear in a sentence. Therefore, a review could be split into a sequence of non-sentences, which are short passages that contain one aspect, and then a response could be generated for each passage. This will enable us to mention the aspects more thoroughly. Finally, the response generation model can be replaced with a large language model such as ChatGPT.

A few ethical considerations should be taken into account. Since a response generation model is trained from reviews and responses on actual hotel booking websites, private information, especially named entities such as the names of people and hotels, might be generated. Furthermore, the use of our system for impersonating a hotel manager may be perceived as inappropriate by customers. Our method can be applicable as a not fully automatic system but a support system that helps hotel managers,where a manual check of the generated responses is necessary to ensure privacy.

[6] The original Japanese texts are shown in Appendix D.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cuiyun Gao, Jichuan Zeng, Xin Xia, David Lo, Michael R. Lyu, and Irwin King. 2019. Automating app review response generation. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 163–175.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*, pages 1275–1312.

Hisatoshi Igusa and Fujio Toriumi. 2021. Automating review response generation using review characteristics (in Japanese). *Proceedings of the 35th Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2021:2F3GS10g01.

Tannon Kew, Michael Amsler, and Sarah Ebling. 2020. Benchmarking automated review response generation for the hospitality domain. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pages 43–52, Barcelona, Spain. Association for Computational Linguistics.

Tannon Kew and Martin Volk. 2022. Improving specificity in review response generation with data-driven data filtering. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 121–133, Dublin, Ireland. Association for Computational Linguistics.

Language Media Processing Lab at Kyoto University. 2021. Japanese BART base. https://huggingface.co/ku-nlp/bart-base-japanese. (accessed May 2024).

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation mo dels. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for C omputational Linguistics*, pages 311–318. Association for Computational Linguistics.

Rakuten Institute of Technology. 2016. RAKUTEN DATA RELEASE. https://rit.rakuten.com/data_release/. (accessed May 2024).

Kalyani Roy, Avani Goel, and Pawan Goyal. 2022. Effectiveness of data augmentation to identify relevant reviews for product question answering. In *Companion Proceedings of the Web Conference*, pages 298–301.

Tohoku NLP Group. 2019. BERT base Japanese (IPA dictionary, whole word masking enabled) – Hugging Face. https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking/. (accessed May 2024).

Lujun Zhao, Kaisong Song, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2019. Review response generation in e-commerce platforms with external product information. In *The World Wide Web Conference*, pages 2425–2435.

A Examples of Sentences Removed by Generality Filtering

Table 10 shows the original Japanese of the sentences in Table 2.

Sentence	Score
大変申し訳ございませんでした。	78544
心よりお詫び申し上げます。	49978
この度は、ご宿泊頂きまして誠に有難うございます。	48500
またのご来館を心よりお待ち申し上げております。	39078
お客様のご指摘はごもっともと受け止めております。	34997

Table 10: Example of the removed general sentences.

B Statistics of Dataset

Table 11 shows the average (ave.) and standard deviation (sd.) of the number of sentences per review/response and the number of words per sentence in the dataset in Table 4.

	review		response	
	ave.	sd.	ave.	sd.
Num. of sentence	5.43	4.38	6.49	2.53
Num. of word	20.9	15.5	20.0	12.6

Table 11: Statistics of the dataset for response generation.

C Instruction to Human Subjects

The detailed instructions to evaluate the generated responses are shown below.

Fluency Rate how natural a response is as a Japanese text on a five-point scale.

1. A considerable number of grammatical errors are found.
2. A few grammatical errors are found.
3. There is no grammatical error, but it is somewhat unnatural.
4. It is an almost natural sentence.
5. It is a completely natural sentence.

Non-redundancy Rate how redundant a response is on a five-point scale.

1. Almost the same sentences are repeated many times.
2. Almost the same sentences are repeated.
3. Almost the same expressions are repeated many times, although their meanings are different.
4. Almost the same expressions are repeated, although their meanings are different.
5. No repetition of the same expressions and sentences is found.

Overall Score Supposing you had written the complaints in the review, how would you feel about the hotel's response? Rate the overall score of it on a five-point scale.

1. Obviously inappropriate.
2. Inappropriate.
3. Neither appropriate nor inappropriate.
4. Appropriate.
5. Obviously appropriate.

Mention of aspect For each aspect in a review, check whether a response mentions the aspect. (All aspects in a review are manually extracted and presented to the evaluator.)

D Example of Generated Responses

Table 12 shows the original Japanese review and generated responses of ones in Table 9.

Review	Response		
	BASELINE-S	PRO-AG-S	GOLD
まず最初に、全室禁煙なら大々的に謳って欲しい。サイトのどこかには記載があるだろうけど、ざっくり読む分には見つけきらない。係りの方は雰囲気の良い方で、満足です。客室のデザインは良いが、お風呂がない。貸切風呂（一番左奥）は温度調整できず真夏に熱い風呂と罰ゲームでした。シャワーが弱すぎる。最終的には湯船の熱いお湯を汲み利用しました。大浴場は普通で良かった。食事は素材の味を生かし、的な内容でした。悪く言えばそんなに手をかけていない。っと思いきや、唯一一品だけ、大根のレモン煮（？）抜群に美味かった。これだけは、感動しました。	貸切風呂の温度につきましては、ご迷惑をお掛け致しまして申し訳ございませんでした。お客様に快適にお過ごし頂けるよう、スタッフ一同努力して参ります。この度は当ホテルをご利用頂きまして誠にありがとうございます。	禁煙ルームのご案内につきましては、ご不便をお掛け致しました事、心よりお詫び申し上げます。ご指摘いただきましたサイトの件でございますが、お客様にご迷惑をお掛け致しましたことを深くお詫び申し上げます。貸切風呂の温度調整につきましては、お客様にご迷惑をお掛け致しました事を心よりお詫び申し上げます。シャワーの件では大変申し訳ございませんでした。今後このような事がないよう、スタッフ一同精進して参ります。	湯の宿　XXXでございます。先日はご宿泊頂きまして、誠にありがとうございました。ご指摘頂きました箇所に関しましては、出来る箇所から順次、改善してまいりたいと思います。今後とも、XXXを宜しくお願い致します。 (XXXはホテル名)

Table 12: Examples of generated responses.