Generation of Dialogue Breakdown Repair Utterance in Non-task Oriented Conversation

Kazuya Tsubokura, Yurie Iribe
Aichi Prefectural University
Nagakute city, Aichi, Japan
{tsubokura, iribe}@ist.aichi-pu.ac.jp

Norihide Kitaoka
Toyohashi University of Technology
Toyohashi city, Aichi, Japan
kitaoka@tut.jp

Abstract

Recent advances in large language models (LLMs) have significantly improved the response quality of dialogue systems. However, the issue of dialogue breakdowns where systems produce utterances that confuse users, such as those containing misinformation or lacking common sense still persists. Since such breakdowns can negatively affect users' impressions of dialogue systems, it is essential to appropriately repair the flow of conversation. Nevertheless, no existing method can robustly handle a wide range of breakdown types. To address this issue, this study aims to develop a dialogue breakdown repair generation system that can robustly handle various types of dialogue breakdowns. Specifically, we train an LLM using the Dialogue Breakdown Repair Corpus, which contains repair utterances corresponding to diverse breakdown scenarios. As a result, we construct a model specialized in generating repair utterances for various breakdown types and demonstrate that it achieves higher accuracy than existing models.

1 Introduction

In recent years, dialogue systems have been increasingly utilized across various industries and domains, such as call centers and healthcare. With the advent of large language models (LLMs), dialogue systems have become capable of producing fluent and natural conversations. However, the issue of dialogue breakdowns (Martinovsky and Traum, 2003) remains unavoidable. Dialogue breakdowns can lead to user frustration (Chakrabarti and Luger, 2015) and may cause confusion (Bickmore et al., 2018). Therefore, considering the possibility of dialogue breakdowns, detecting them and repairing the conversational flow is crucial for maintaining smooth interactions.

An example of applying dialogue breakdown detection to dialogue systems is the work by Inaba and Takahashi (Inaba and Takahashi, 2019). In this study, dialogue breakdown detection is applied to response candidates before the system generates an utterance. By reranking these candidates and prioritizing those with a lower likelihood of causing breakdowns, the system's response performance is improved. However, even when selecting responses with a low probability of causing breakdowns, it is difficult to eliminate them entirely. Thus, methods for repairing the conversational flow after a breakdown are necessary.

Several approaches for repairing dialogue breakdowns have been proposed. Uchida et al. (Uchida et al., 2019) proposed a strategy that recovers from dialogue breakdowns by encouraging user cooperation, making users aware that they share responsibility for the breakdown. This method classifies errors into four types (listening errors, understanding errors, automatic speech recognition errors, and interruption errors) and adapts recovery strategies accordingly. However, the types of errors it can handle are limited, and since it relies on rule-based repair, it lacks generalizability and cannot handle undefined error types.

Sugiyama et al. (Sugiyama et al., 2018) proposed a framework in which two robots collaboratively engage in dialogue to sustain long conversations while avoiding breakdowns. They demonstrated that cooperative behaviors such as appropriate switching of speaker robots and inter-robot question-answer interactions significantly reduce the perception of dialogue breakdowns. However, this approach employs a fixed dialogue flow for robot collaboration, which prevents users from interrupting the robots' utterances.

As described above, while several methods for repairing dialogue breakdowns have been explored, challenges remain in terms of generalizability. Therefore, there is a need for repair strategies that can robustly handle a wide range of potential errors.

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 553–558
August 2–5, 2026.

Table 1: Example of a dialogue breakdown and crowd worker responses (Originally in Japanese; translated into English)

Breakdown example presented to crowd workers
User: Have you ever been to Tokyo Tower? System (breakdown utterance): I have never been there. It is a 555-meter-tall tower, right?
Crowd worker responses
User (response to the breakdown): Was it really that tall? System (repair utterance): Sorry, I made a mistake. It is 333 meters tall.

In this study, we aim to develop a dialogue breakdown repair generation system that can robustly handle various types of dialogue breakdowns. Specifically, we train an LLM using the Dialogue Breakdown Repair Corpus (Tsubokura et al., 2026), which contains system utterances designed to repair different types of dialogue breakdowns. Through this approach, we construct a model specialized in generating repair utterances that can address a wide variety of dialogue breakdown types.

2 Dialogue Breakdown Repair Corpus

The Dialogue Breakdown Repair Corpus (DBR Corpus) (Tsubokura et al., 2026)[1] is a Japanese corpus that collects examples of how a system should repair its responses when a breakdown occurs in open-domain conversational dialogue. In the DBR Corpus, crowd workers are presented with dialogue histories leading up to a system breakdown. From these, two types of utterances are collected: (1) user responses to the system's breakdown utterance, and (2) system responses expected by the user to repair the dialogue in reaction to the user's utterance.

Following the taxonomy of dialogue breakdowns proposed by Higashinaka et al. (Higashinaka et al., 2022), the authors constructed 140 breakdown scenarios and collected 3,990 responses from 57 crowd workers. Therefore, the DBR Corpus can be regarded as a comprehensive and rich linguistic resource that extensively covers various types of dialogue breakdowns that may occur in text-based open-domain dialogue systems.

[1] https://github.com/kzy-tbkr/Dialogue-Breakdown-Repair-Corpus

3 Generation of Repair Utterances

3.1 Baseline Models

In this section, we evaluate zero-shot inference performance for both a closed model and several open models, and compare their performance. As the closed model, we use GPT-4o-mini (gpt-4o-mini-2024-07-18). As open models, we use llm-jp-3-3.7b-instruct[2] and DeepSeek-R1-Distill-Qwen-7B-Japanese[3].

For each model, we provide the prompt shown in Table 7 in Appendix A, and then generate repair utterances.

Since whether to perform a repair is considered to depend on the dialogue context and the interlocutor's individual characteristics, we do not explicitly instruct the model to perform a repair in the prompt. Instead, the model is asked to generate "an appropriate response for the next utterance."

3.1.1 Evaluation Method

For each selected model, repair utterances were generated and evaluated. Specifically, repair utterances were generated for 399 randomly sampled test instances, and their similarity to the utterances provided by crowd workers was evaluated.

The evaluation metrics are listed below. For all metrics, higher values indicate greater similarity.

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is one of the standard metrics used to evaluate machine translation quality. In this study, the repair utterances provided by crowd workers are treated as references, and the similarity between these references and the utterances generated by the proposed method is computed.

BERTScore (Zhang et al., 2020) is a similarity metric based on contextual embeddings. It has been reported to correlate well with human judgments. We use *xlm-roberta-large*[4] (Conneau et al., 2019) provided by Facebook AI as the underlying BERT model.

Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019) is a method for generating sentence embeddings. By representing utterances as vectors, we compute the cosine similarity between the repair utterances provided by crowd workers and those generated by the proposed

[2] https://huggingface.co/llm-jp/llm-jp-3-3.7b-instruct

[3] https://huggingface.co/lightblue/DeepSeek-R1-Distill-Qwen-7B-Japanese

[4] https://huggingface.co/FacebookAI/xlm-roberta-large

Table 2: Evaluation results of zero-shot repair utterance generation with baseline models (mean values are shown, with standard deviations in parentheses)

	BLEU	BERTScore	S-BERT
gpt-4o-mini	**6.7692** (6.9579)	**0.8889** (0.0222)	**0.4795** (0.2228)
llm-jp3.7b	5.8347 (9.1477)	0.8740 (0.0387)	0.4034 (0.2301)
DeepSeek-7b	0.8297 (0.0190)	0.8487 (0.8679)	0.2956 (0.1858)

method to measure their similarity. We use *paraphrase-multilingual-MiniLM-L12-v2*[5] as the model.

3.1.2 Experimental Results

The experimental results are shown in Table 2. As can be seen from the table, GPT-4o-mini achieves the highest performance across all three metrics. Although details such as the number of parameters and training data of GPT-4o-mini are not publicly available, it outperforms the two open models.

3.2 Performance Evaluation of Fine-Tuning

In this section, we compare the repair utterance generation performance between a pre-existing model and a model fine-tuned using the collected data, in order to evaluate the repair capabilities of general-purpose models that are not specifically optimized for repair utterance generation.

In this experiment, we use GPT-4o-mini, which achieved the best performance in the previous section, for repair utterance generation. Specifically, as the pre-existing model, we prompt the model with dialogue contexts containing system breakdowns and generate the next appropriate response. In addition, we generate utterances using a model fine-tuned on the repair utterances expected by crowd workers.

We fine-tuned GPT-4o-mini using the collected dataset. The dataset, consisting of 3,990 instances, was split into training, validation, and test sets with a ratio of 8:1:1. The model was trained using the training and validation sets. The hyperparameters were set as follows: n_epochs = 3 and learning_rate_multiplier = 1.0.

3.2.1 Evaluation Method

Using the same evaluation metrics as in the previous experiment, we compare generation by the pre-existing model (Prompting) and the fine-tuned model (Fine-tuning). Repair utterances are generated for 399 test instances, and their similarity to the utterances provided by crowd workers is evaluated.

[5] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Table 3: Evaluation results of repair utterance generation using GPT-4o-mini with zero-shot prompting (Prompting) and fine-tuning (Fine-tuning).

	BLEU	BERTScore	S-BERT
Prompting	6.769	0.889	0.480
Fine-tuning	**12.843**	**0.914**	**0.522**

Table 4: Average number of sentences and morphemes for repair utterances generated by the crowd workers, GPT-4o-mini with zero-shot prompting (Prompting), and GPT-4o-mini with fine-tuning (Fine-tuning).

	Crowd workers	Prompting	Fine-tuning
Ave. # of sentences	1.33	2.66	1.22
Ave. # of morphemes	10.57	27.68	9.18

3.2.2 Experimental Results

The evaluation results are shown in Table 3. The fine-tuned model achieves higher similarity to the utterances provided by crowd workers across all three metrics.

3.2.3 Discussion

The experimental results show that the fine-tuned GPT-4o-mini achieves higher similarity to the utterances provided by crowd workers. This suggests that fine-tuning with the collected dataset enables the model to generate repair utterances that better align with human expectations.

Closed models tend to generate polite and lengthy responses, which can be redundant or overly formal in casual conversations. Therefore, we conduct a more detailed analysis in the following section.

Sentence and Morpheme Counts First, we focus on the length of the generated utterances and compute the number of sentences and morphemes in the repair utterances. The results are shown in Table 4.

To compare the number of sentences and morphemes among the three methods (crowd worker

Table 5: LLM-based evaluation results for repair utterances generated by the crowd workers, GPT-4o-mini with zero-shot prompting (Prompting), and GPT-4o-mini with fine-tuning (Fine-tuning).

	Crowd workers	Prompting	Fine-tuning
Redundancy	1.93	2.79	1.82
Politeness	3.13	3.63	3.24
Informativeness	2.23	3.15	2.13
Naturalness	3.93	4.09	4.03

responses, Prompting, and Fine-tuning), we conducted a Friedman test. The results showed a significant difference (the number of sentences: $\chi^2(2) = 617.9, p < 0.001$, the number of morphemes: $\chi^2(2) = 568.2, p < 0.001$). Therefore, we further conducted a Wilcoxon signed-rank test as a post-hoc analysis. The results indicate that significant differences were observed between all pairs of methods ($p < 0.001$). The significance level was set at $\alpha = 0.05$. No correction for multiple comparisons was applied.

These results suggest that the Prompting method without fine-tuning tends to generate relatively longer utterances, whereas the Fine-tuning method produces shorter outputs, similar to those provided by crowd workers.

LLM-based Evaluation Based on the LLM-as-a-Judge paradigm, we evaluate the generated utterances using four criteria: verbosity, politeness, informativeness, and naturalness. We use GPT-4o as the evaluation model. The prompt used for evaluation is shown in Table 8 in Appendix A.

The evaluation results are presented in Table 5. To compare the evaluation scores among the three methods (crowd worker responses, Prompting, and Fine-tuning), we conducted a Friedman test. The results showed a significant difference (redundancy: $\chi^2(2) = 385.0, p < 0.001$, politeness: $\chi^2(2) = 204.5, p < 0.001$, informativeness: $\chi^2(2) = 492.1, p < 0.001$, naturalness: $\chi^2(2) = 36.5, p < 0.001$), and therefore we performed a Wilcoxon signed-rank test as a post-hoc analysis. The results indicate that significant differences were observed between all pairs of methods ($p < 0.05$). The significance level was set at $\alpha = 0.05$. No correction for multiple comparisons was applied.

Prompting achieves the highest scores across all four criteria: redundancy, politeness, informativeness, and naturalness. In particular, it shows higher scores in redundancy and informativeness compared to crowd worker responses and Fine-tuning, indicating that utterances generated by Prompting tend to be more verbose, polite, and informative.

While higher politeness and informativeness are generally desirable for dialogue systems, this study focuses on repair in casual conversations, where excessively polite and information-rich responses are not necessarily appropriate. Since it is desirable for generated utterances to resemble those produced by crowd workers, the utterances generated by Prompting can be considered overly verbose and informative, and thus not fully suitable for repair in casual dialogue.

In contrast, utterances generated by Fine-tuning exhibit scores closer to those of crowd worker responses. This suggests that fine-tuning the model with the collected dataset enables the generation of repair utterances that are more human-like.

Examples of Generated Utterances Examples example of utterances generated by the two methods is shown in Table 6. Comparing the two methods, the utterances generated by Prompting tend to be longer and more verbose, whereas those generated by Fine-tuning are shorter and closer to the responses provided by crowd workers.

4 Conclusion and Limitations

To address the problem of dialogue breakdowns in dialogue systems, this study fine-tunes an LLM using the Dialogue Breakdown Repair Corpus. Compared to existing LLMs, the model fine-tuned on the proposed corpus generates utterances that are closer to the ground-truth responses provided by crowd workers across four dimensions: redundancy, politeness, informativeness, and naturalness.

Since this study is limited to fine-tuning, future work will leverage additional information in the DBR Corpus, such as the degree of breakdown and individual user traits, to further improve the generation of more appropriate repair utterances.

As a limitation of this study, the dataset is restricted to Japanese. Since repair strategies and user preferences may vary across languages and cultures, it is necessary to investigate whether the proposed approach can be generalized to languages other than Japanese.

Furthermore, this study relies solely on automatic evaluation. Therefore, human evaluation

Table 6: Examples of repair utterances generated by the crowd workers, GPT-4o-mini with zero-shot prompting (Prompting), and GPT-4o-mini with fine-tuning (Fine-tuning) (Originally in Japanese; translated into English).

	Example 1
Dialogue History	User: It's so hot today, I feel like eating ice cream. / System: Ice cream sounds great! Want to go buy some? / User: Yeah, let's go to a convenience store. / System: Aren't you going to the convenience store? / User: I am, who said I'm not?
Crowd workers	Let's go, then.
Prompting	Sorry, I misunderstood! What kind of ice cream are you planning to get?
Fine-tuning	Oh, you're going. That's good.
	Example 2
Dialogue History	User: What is your favorite food? / System: Cheesecake. / User: It's delicious, I like it too. / System: Yes, I like cheesecake. / User: You really seem to like cheesecake.
Crowd workers	Yes, it's my favorite.
Prompting	Yes, it's creamy and perfectly sweet, so I can't resist it. What kind of cheesecake do you like?
Fine-tuning	Yes, I love cheesecake.
	Example 3
Dialogue History	User: What do you usually do on your days off? / System: Recently, I've been into watching soccer, so I often watch it on my days off. / User: I like soccer too. Do you have a favorite team? / System: Yes. / User: Which team do you support?
Crowd workers	I support Vissel Kobe.
Prompting	I support FC Barcelona. Which team do you support?
Fine-tuning	I support FC Tokyo.

in real interactive settings between users and dialogue systems is required for a more comprehensive assessment.

Acknowledgments

This work is supported by Adaptable and Seamless Technology transfer Program through Target-driven R&D (A-STEP) from Japan Science and Technology Agency (JST) Japan Grant Number JPMJTR25R5. This work was supported by JSPS KAKENHI Grant Number JP23H00493.

References

Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. 2018. *Safety First: Conversational Agents for Health Care*, pages 33–57. Springer International Publishing, Cham.

Chayan Chakrabarti and George F. Luger. 2015. Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications*, 42(20):6878–6897.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2022. Classification of utterances that lead to dialogue breakdowns in chat-oriented dialogue systems (in japanese). *Journal of Natural Language Processing*, 29(2):443–466.

Michimasa Inaba and Kenichi Takahashi. 2019. Applying dialogue breakdown detection to dialogue systems (in japanese). *Transactions of the Japanese Society for Artificial Intelligence*, 34(3):B–I64_1–8.

Bilyana Martinovsky and David Traum. 2003. The error is the clue: breakdown in human-machine interaction. In *Error Handling in Spoken Dialogue Systems*, pages 11–16.

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 Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hiroaki Sugiyama, Toyomi Meguro, Yuichiro Yoshikawa, and Junji Yamato. 2018. Avoiding breakdown of conversational dialogue through inter-robot coordination. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 2256–2258, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Kazuya Tsubokura, Yurie Iribe, and Norihide Kitaoka. 2026. A corpus for personalized dialogue breakdown repair in japanese open-domain conversations. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2899–2912, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Takahisa Uchida, Takashi Minato, Tora Koyama, and Hiroshi Ishiguro. 2019. Who is responsible for a dialogue breakdown? an error recovery strategy that promotes cooperative intentions from humans by mutual attribution of responsibility in human-robot dialogues. *Frontiers in Robotics and AI*, 6.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompt

This appendix presents the prompts used in our experiments. Table 7 shows the prompt used for repair utterance generation, and Table 8 shows the prompt used for evaluation. The "Dialogue History" in the prompt consists of the dialogue breakdown examples presented to crowd workers, along with the users' response utterances after the system breakdown provided by the crowd workers.

Table 7: Prompt for repair utterance generation (Originally in Japanese; translated into English)

You are a dialogue system. Based on the dialogue context, generate an appropriate response for the next utterance. {Dialogue History}

Table 8: Prompt for evaluating repair utterances (Originally in Japanese; translated into English)

You are a conversation analyst. For the given dialogue context, evaluate the "Target: the final system utterance" using the following four criteria, and return the results in JSON format. **[Criteria and Scales]** - Redundancy (redundancy_score: integer from 1 to 5) 1 = very concise, 2 = concise, 3 = neutral, 4 = somewhat redundant, 5 = redundant - Politeness (politeness_score: integer from 1 to 5) 1 = too casual, 2 = somewhat casual, 3 = neutral, 4 = polite, 5 = very polite - Informativeness (informativeness_score: integer from 1 to 5) 1 = no information, 2 = little information, 3 = moderate, 4 = sufficient, 5 = rich and specific - Naturalness in Japanese conversation (naturalness_score: integer from 1 to 5) 1 = unnatural, 2 = somewhat unnatural, 3 = neutral, 4 = natural, 5 = very natural **[Instructions]** - Consider the dialogue context and evaluate only the final system utterance. - The output must be in the following JSON format only: { "redundancy_score": $< 1-5 >$, "politeness_score": $< 1-5 >$, "informativeness_score": $< 1-5 >$, "naturalness_score": $< 1-5 >$, "label": "brief summary label (e.g., concise and polite / somewhat redundant but informative)", "rationale": "1–2 sentence justification" }