I Remember You!: SUI Corpus for Remembering and Utilizing Users' Information in Chat-oriented Dialogue Systems

Yuiko Tsunomori[1], **Ryuichiro Higashinaka**[1]
[1]Graduate School of Informatics, Nagoya University, Japan
{tsunomori.yuiko.u9@s.mail, higashinaka@i}.nagoya-u.ac.jp

Abstract

To construct a chat-oriented dialogue system that will be used for a long time by users, it is important to build a good relationship between the user and the system. To achieve a good relationship, several methods for remembering and utilizing information on users (preferences, experiences, jobs, etc.) in system utterances have been investigated. One way to do this is to utilize user information to fill in utterance templates for use in response generation, but the utterances do not always fit the context. Another way is to use neural-based generation, but in current methods, user information can be incorporated only when the current dialogue topic is similar to that of the user information. This paper tackled these problems by constructing a novel corpus to incorporate arbitrary user information into system utterances regardless of the current dialogue topic while retaining appropriateness for the context. We then fine-tuned a model for generating system utterances using the constructed corpus. The result of a subjective evaluation demonstrated the effectiveness of our model. Furthermore, we incorporated our fine-tuned model into a dialogue system and confirmed the effectiveness of the system through interactive dialogues with users.

Keywords: Dialogue corpus, Chat-oriented dialogue system, User information

1. Introduction

The demand for chat-oriented dialogue systems has been increasing in both research and commercial fields (Adiwardana et al., 2020; Shuster et al., 2022). To construct a chat-oriented dialogue system that will be used for a long time by users, it is important to build a good relationship between the user and the system, which requires that the user and the system know each other well (Richards and Bransky, 2014; Bickmore and Picard, 2005). In human-to-human dialogue, it is effective to remember and utilize information on the dialogue partner, such as preferences and experiences disclosed by the other party, for building a good relationship (Hall, 2019). To build a good relationship between the system and the user, several methods that remember and utilize information on users (called "**user information**" in this paper) in system utterances have been investigated. Tsunomori et al. (2019) constructed a chat-oriented dialogue system that remembers and utilizes user information obtained from past dialogue and experimentally confirmed that incorporating user information into system utterances improves users' familiarity with chat-oriented dialogue systems. However, in their work, the system utterances were generated using templates to be filled with user information, which often caused inappropriate utterances with regard to the context. Xu et al. (2022b) used neural-based models with dialogue context and user information as input to generate system utterances. However, in their method, user information can be incorporated only when the current dialogue topic is similar to that of the user information. This limits the opportunities for systems to utilize user information because user information similar to the current dialogue topic may not always be available in real-world settings.

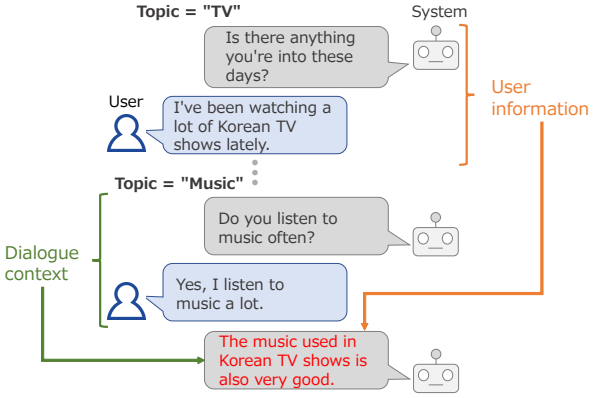

Figure 1: Example system utterance based on user information and dialogue context.

We aim to realize a personalized chat-oriented dialogue system that builds a good relationship with users by remembering and utilizing arbitrary user information naturally and actively. Figure 1 shows a dialogue example from the personalized chat-oriented dialogue system that we aim to achieve. In the figure, the system references the user information extracted from its previous interaction with the user and then weaves it into its utterance.

In this paper, to realize such a personalized chat-oriented dialogue system, we constructed

LREC-COLING 2024, pages 9285–9295
20-25 May, 2024.

the **S**ystem utterance based on **U**ser **I**nformation corpus (**SUI corpus**) by extending an existing dialogue corpus. The language of the corpus is Japanese. The SUI corpus contains triplets formed of ⟨user information, dialogue context, system utterance based on the user information and dialogue context (**expanded system utterance**)⟩. With this corpus as a basis, we constructed a model for generating system utterances. Our contributions are as follows.

- We constructed the SUI corpus; this is a novel corpus consisting of utterances incorporating various kinds of user information regardless of the current dialogue topic. The SUI corpus is publicly available.[1]
- We fine-tuned a model to generate system utterances using the SUI corpus and conducted a subjective evaluation. The results showed that our model could incorporate arbitrary user information into system utterances regardless of the current dialogue topic while retaining appropriateness for the context.
- We incorporated our fine-tuned model into a dialogue system and confirmed the effectiveness of the system through a live interactive evaluation.

2. Related Work

There are several studies on personalizing system utterances in chat-oriented dialogue systems by utilizing user information extracted from dialogues with heuristic rules. Sugo and Hagiwara (2014) used rules to extract user information and used them in system utterances to show the user that the system can remember user information. Their system selects its utterances on the basis of the acquired preferences of the user. Tsunomori et al. (2019) constructed a chat-oriented dialogue system that extracts and uses user information and confirmed the effectiveness of the system through evaluations of interactive dialogue with users. They reported that remembering and utilizing user information were important to make users feel familiar with a dialogue system. These studies use rules and templates for system utterance generation, which often makes it difficult to generate utterances while retaining appropriateness to the context.

Recently, neural-based methods for utilizing user information for system utterance generation have been proposed. These methods can generate more natural utterances on the basis of dialogue contexts. Xu et al. (2022a) constructed a dialogue model that creates user information summaries from dialogue histories and uses them as the dialogue contexts for utterance generation. However, they did not incorporate arbitrary user information into system utterances. Similarly, Xu et al. (2022b) constructed a neural-based chat-oriented dialogue system that incorporates user information into system utterances. The model selects stored user information close to the current dialogue topic and uses the user information and dialogue context as input to generate system utterances. While it is reasonable to bring up user information related to previous system utterances when the topic is similar, we consider this to severely limit opportunities for the system to utilize user information. Therefore, this paper focuses on utterance generation using user information regardless of the current dialogue topic. It is also worth noting that Xu et al. (2022b) only performed evaluations using a user simulator; it is not known if the model will work effectively in interactive dialogue systems with users. We verify our model's effectiveness by incorporating the model into dialogue systems and evaluating the systems with users through a live interactive evaluation.

3. System Utterance Based on User Information Corpus (SUI Corpus)

To achieve utterance generation that utilizes arbitrary user information while retaining appropriateness for the context, we constructed the SUI corpus on the basis of dialogue contexts and user information.

3.1. Overview

The SUI corpus was constructed by extending the existing Osaka University Multimodal Dialogue Corpus (Hazumi) (Komatani et al., 2019). The Hazumi corpus is a person-to-system multimodal corpus in Japanese consisting of spoken dialogues between users and systems operating under the Wizard-of-Oz (WoZ) method. The wizard selects the system's responses through a dedicated interface and changes the topic in accordance with the user's interests when selecting utterances in the dialogues. For each of the dialogues, the wizard does not repeat the same topics. We used speech transcriptions from Hazumi1911, which consists of 30 dialogues by 30 users (2,859 turns in total). We chose this corpus because it contains dialogues in which a system talks about topics related to the user's interests, so we assumed it would contain a lot of user information.

Figure 2 shows the flow for constructing the SUI corpus, which consisted of two tasks: (1)

[1] https://github.com/nu-dialogue/sui-corpus

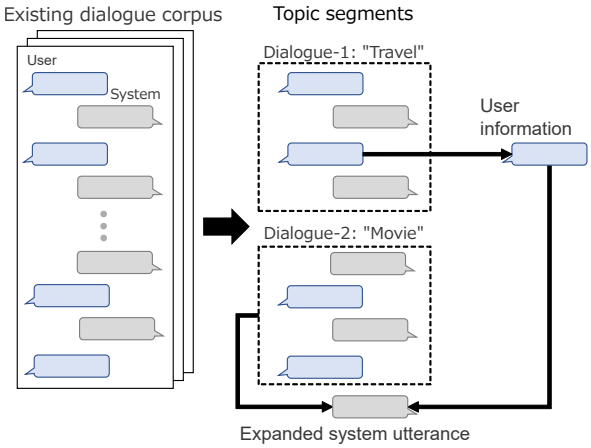

Figure 2: Flow of SUI corpus construction.

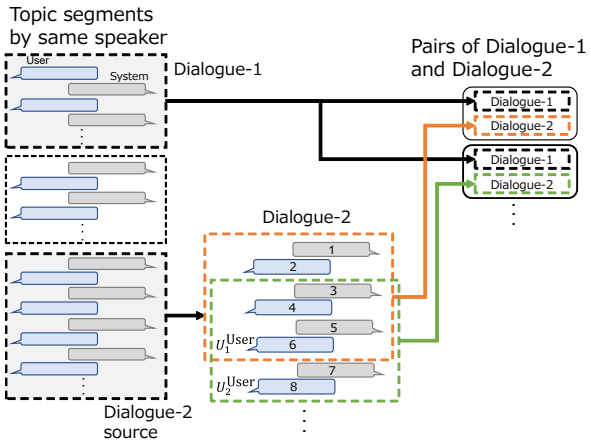

Figure 3: Procedure for creating pairs of dialogue-1 and dialogue-2.

extraction of user information from a dialogue (called **dialogue-1**) and (2) creation of system utterances based on the user information extracted in task (1) and another dialogue (called **dialogue-2**). Dialogue-1 and dialogue-2 are dialogues in which the same user talks about different topics. We first divided each dialogue in Hazumi into topic segments and then created pairs of dialogue-1 and dialogue-2. Then, we collected expanded system utterances based on the pairs. In the following subsections, we provide details on the creation of the dialogue pairs and the two tasks.

3.2. Pairs of Dialogue-1 and Dialogue-2

We created pairs of dialogue-1 and dialogue-2 (which are on different topics) to collect expanded system utterances. Dialogue-1 was used to extract user information, and dialogue-2 was used for dialogue context. First, we removed fillers and misspellings from the Hazumi1911 transcriptions by using heuristic rules, and then we divided the dialogues into topic segments using fixed utterances ("Let's talk about [topic word]!", "Now, I would like to move on to the next topic," and "So, this is the last question.") as delimiters that the wizard used to change topics. Approximately five topics were discussed per dialogue. We removed short topic segments with less than 15 utterances and created a total of 152 topic segments.

Figure 3 illustrates the procedure for creating pairs of dialogue-1 and dialogue-2, which is described as follows.

1. We select two dialogues from different topic segments by the same speaker. In chronological order, the one spoken earlier is dialogue-1, and the latter is a **dialogue-2 source**.
2. We extract a portion of the dialogue-2 source to create dialogue-2. Specifically, let U_1^{user} be the first user utterance that appears after the sixth utterance in the dialogue-2 source. From U_1^{user}, we extract a total of six utterances going back in time and name that portion of utterances dialogue-2. We extract dialogue-2 from U_2^{user} in the same way. This is repeated until we reach U_N^{user}, where N is the index of the last user utterance in the dialogue-2 source.
3. Repeat 1–2 for all topic segments by the same speaker.

After completing the above procedure for the data of all speakers, we had a total of 1,594 pairs of dialogue-1 and dialogue-2. Note that "dialogue-2" and "dialogue-2 source" are different entities. Dialogue-2 source refers to a topic segment obtained by dividing a dialogue in the Hazumi corpus by dialogue topics, whereas dialogue-2 is a dialogue obtained by dividing dialogue-2 source into groups of six utterances each.

To confirm whether the SUI corpus consists of user information with varied relevance to the dialogue context, we investigated the similarity between user information and dialogue contexts in the corpus. Specifically, we calculated the similarity of topic words between pairs of dialogue-1 and dialogue-2. We extracted word embeddings of the topic words using FastText (Bojanowski et al., 2017) trained by Wikipedia[2] and then calculated the cosine similarity between them. The similarity score range between 0.2 and 0.3 had the highest frequency. The highest similarity score was 0.48 ("movie" and "music"), and the lowest score was 0.09 ("sport" and "book"). For reference, we also calculated the cosine similarity of Japanese synonyms. We used Japanese WordNet (version 1.1) [3] (Bond et al., 2012) and extracted synonyms belonging to the same synset. We calculated the cosine similarity between all pairs of words in each

[2] https://github.com/Hironsan/awesome-embedding-models

[3] https://bond-lab.github.io/wnja/

synset, and the mean was 0.40. This result indicates that dialogue-1 contains topics with various degrees of similarity to dialogue-2. Using the pairs, we can expect to collect expanded system utterances incorporating various user information with a wide degree of relevance to the dialogue context.

3.3. Collecting Expanded System Utterances

We used Lancers[4], a crowdsourcing service in Japan, to collect expanded system utterances for the SUI corpus. The workers created seven utterances for each pair of dialogue-1 and dialogue-2 in the following steps. The number of utterances to create was set to seven in consideration of the load on workers.

1. From dialogue-1, as user information, extract as many user utterances that contain self-disclosures as possible. If the user utterance alone is not self-contained, the previous system utterance should also be extracted as part of the user information. For example, consider a situation where the system asks, "What is your favorite food?" and the user answers, "Apple." In this case, since the user utterance alone is insufficient as self-disclosure, the previous system utterance should also be extracted.
2. Select seven of the extracted user information items. If there are less than seven, select a total of seven overlapping items that can be used to create expanded system utterances.
3. For each of the user information items selected in step 2, create an expanded system utterance by considering both the user information and dialogue-2 (dialogue context). When using the same user information, create different utterances. Note that it is not allowed to create utterances that forcibly incorporate user information items by using phrases such as "By the way," "Speaking of," and so on.

In total, 34 workers participated, and 10,801 expanded system utterances were collected.

Table 1 shows example data from the SUI corpus, where the user talked about "drinking alcohol" in the past dialogue (user information) and is talking about "listening to classical music" in the current dialogue (dialogue context). The expanded system utterance, "Do you ever enjoy your favorite classical music while drinking alcohol?" naturally associated "listening to classical music" with "drinking alcohol." Table 2 shows the statistics of the SUI corpus. Here, MeCab[5] was used for word segmentation. Compared with Hazumi1911 (original), the expanded system utterances were longer and contained more words, reflecting the fact that the user information was incorporated.

3.4. Quality Assessment

We conducted a quality assessment of the SUI corpus by using CrowdWorks[6], a crowdsourcing service in Japan. We randomly selected 1,000 expanded system utterances, and then each utterance was evaluated by three workers. We presented the workers with the user information, dialogue context, and expanded system utterances for assessment and had them judge each of the following three items on a binary scale of "Yes/No" for each expanded system utterance.

Dialogue context reflection Is the expanded system utterance based on the dialogue context?
User information reflection Is the expanded system utterance based on the user information?
Naturalness Do you feel that the expanded system utterance is natural?

Note that we imposed a binary decision here based on our preliminary study that indicated the difficulty of judging the degree of how much user information is included in an utterance.

Table 3 lists the results of the quality assessment. The annotation agreement statistics (Fleiss' κ) show that the agreement rate exceeded 0.5 for dialogue context reflection and user information reflection, which indicates moderate agreement. In contrast, the agreement rate for naturalness was poor, which suggests that the judgment of naturalness is highly subjective. This is in accordance with previous work that shows that subjective single-turn evaluations in dialogues tend to show low agreement (Higashinaka et al., 2015; Ghandeharioun et al., 2019).

The proportions of "Yes" responses to dialogue context reflection and user information reflection were both over 70%. This indicates that a high proportion of expanded system utterances were based on both user information and dialogue contexts. In contrast, the proportion of "Yes" for naturalness was less than 60%, reflecting the possible difficulty of incorporating user information smoothly even for humans.

We further assessed the quality of the expanded system utterances to ascertain whether they could be good enough references for training generation models even if not very natural. For this assessment, we performed dialogue breakdown annotations and investigated whether the naturalness

[4] https://www.lancers.jp/

[5] https://taku910.github.io/mecab/

[6] https://crowdworks.jp

User information	S	Do you like to drink alcohol? (お酒はお好きですか？)
	U	Yes, I drink beer, sake, shochu, and most other alcohol. (そうですね ビールとか 日本酒 酎ハイ 大概のものは飲みます)
Dialogue context	S	What kind of music do you listen to, Japanese, Western, or classical? (聞くとしたら、邦楽，洋楽，クラシックなどどんなジャンルを聞きますか？)
	U	I often listen to Japanese music, but sometimes I also listen to classical music. (そうですねえ 聴く 邦楽が多いですがたまにクラシックも聴きますね)
	S	What genre of music do you like? (どんなジャンルの曲が好きなんですか？)
	U	Yes, I like pop music. For classical music, I like Tchaikovsky. (そうですね ポップス系が好きですね あとクラシックだと チャイコフスキーなどが好きですね)
	S	You like classical music? Is there a favorite composer or music that you often listen to? (クラシックが好きなんですか。よく聞く「作曲者」やきょくってありますか？)
	U	I like Tchaikovsky's Violin Concerto No. 1 the best. (チャイコフスキーのヴァイオリン協奏曲第一番が一番好きですね)
Original	S	What do you like about that? (そのきょくのどういうところが好きなんですか？)
Expanded	S	Do you ever enjoy your favorite classical music while drinking alcohol? (お好きなクラシックとお酒を、一緒に楽しまれることも多いんですか？)

Table 1: Example of SUI corpus, where "original" is system utterance from Hazumi1911, and "expanded" refers to system utterances collected by crowdsourcing. S and U stand for system and user utterances, respectively. All utterances were originally in Japanese. English translations were done by authors.

	No. of letters	No. of words
Original	20.78	12.05
Expanded	34.86	20.39

Table 2: Statistics of system utterances in SUI corpus. "Original" means system utterances from Hazumi1911. "Expanded" means system utterances we collected by crowdsourcing.

	Fleiss' κ	"Yes" ratio
DC-r	**0.55**	0.77
UInfo-r	0.50	**0.87**
Natural	0.22	0.56

Table 3: Quality assessment results for expanded system utterances. "DC-r" represents dialogue context reflection, "UInfo-r" represents user information reflection, and "Natural" represents naturalness. **Bold** font represents top score.

	Hazumi1911	SUI
NB (Not a breakdown)	**0.73**	**0.47**
PB (Possible breakdown)	0.21	0.38
B (Breakdown)	0.06	0.15

Table 4: Ratio of dialogue breakdown annotations given to each corpus.

of the SUI corpus was acceptable compared with the original corpus (Hazumi1911). Here, dialogue breakdown means a situation in a dialogue where users cannot proceed with the conversation (Martinovsky and Traum, 2006). The following three breakdown labels (Higashinaka et al., 2016) were used to annotate each expanded system utterance:

NB Not a breakdown: It is easy to continue the conversation.

PB Possible breakdown: It is difficult to continue the conversation smoothly.

B Breakdown: It is difficult to continue the conversation at all.

Here, the labels indicate how easy/difficult it is to continue the conversation after the system utterance in question. We recruited three workers via Lancers to evaluate each system utterance. The workers subjectively evaluated 200 utterances, including 100 randomly selected expanded system utterances and 100 randomly selected original system utterances from Hazumi1911.

Table 4 shows the results for the dialogue breakdown annotations. Compared with Hazumi1911, the ratio of NB in the SUI corpus was lower, and that of PB was higher. The ratio of B increased slightly, suggesting that forcibly utilizing user information that has a topic different from the current one increases the number of unnatural utterances to some extent. However, we found the quality of the SUI corpus to be reasonable overall, not causing dialogue breakdowns most of the time ($0.47 + 0.38 = 0.85$).

4. Utterance Generation Experiment

We investigated whether a dialogue model fine-tuned by the SUI corpus generates system utterances that incorporate arbitrary user information into system utterances regardless of the current dialogue topic. To verify the performance of dialogue models, automatic evaluation metrics such as BLEU are commonly used (Liu et al., 2016; Zhang et al., 2020). However, it has been reported that there is little correlation with human evaluation (Liu et al., 2016). Thus, we evaluated the model only through a subjective evaluation.

4.1. Fine-tuning Settings

We fine-tuned an existing pre-trained model using the SUI corpus. We used an encoder-decoder model based on Transformer (Adiwardana et al., 2020; Roller et al., 2021) as a pre-trained model. Specifically, we used a Japanese Transformer encoder-decoder dialogue model trained with a large amount of Twitter reply pairs and the Japanese PersonaChat corpus (Sugiyama et al., 2023) as the pre-trained model. The number of parameters is 1.6B. The SUI corpus was divided into train/dev/test datasets. We split the 30 dialogues in Hazumi1911 into train: dev: test $= 24: 3: 3$ by avoiding overlapping dialogues by the same speaker.

SentencePiece (Kudo and Richardson, 2018) was used for tokenization. We used an NVIDIA Tesla V100 as the GPU. As the hyperparameters used in training, the batch size was 8, the optimizer was Adam, and the loss function was a label-smoothed cross-entropy. The learning rate was 1e-04 with a minimum learning rate of 1e-09. The learning rate schedule used inverse sqrt. We applied an early stopping strategy with a patience of 5 and evaluated the model on the dev data at each epoch. The model with the lowest validation loss was utilized for the evaluation.

4.2. Evaluation Settings

The two models we compared are as follows.

Fine-tuned (ours) A model was fine-tuned with the SUI corpus. The input was dialogue context and user information. The input format was "tokenized user information [SEP] tokenized dialogue context."

Vanilla A vanilla Japanese Transformer encoder-decoder dialogue model, not fine-tuned with the SUI corpus. The input was only dialogue context.

We used Lancers to evaluate the generated system utterances. We created an evaluation dataset that included a total of 200 utterances broken down into 100 system utterances generated by each of the two models for the same input. Three workers evaluated each utterance. The other evaluation settings were the same as those for the quality assessment described in Section 3.4.

4.3. Results and Analysis

Table 5 lists the results of the evaluation. We used the Wilcoxon signed-rank test (Wilcoxon, 1945) for the statistical test. The fine-tuned model had a higher average score than the vanilla model, especially for the user information reflection score. These results indicate that these models can effectively generate utterances on the basis of user information and dialogue context. The naturalness score for the fine-tuned model was slightly lower than for the vanilla model, but we believe that this is acceptable because the model was forced to incorporate new information.

	DC-r	UInfo-r	Natural	Ave.
a. Vanilla	**0.81**	0.27	**0.83**[b]	0.64
b. Fine-tuned	0.76	**0.87**[a]	0.71	**0.78**

Table 5: Results of subjective evaluation. Percentage of "Yes" for each item is shown. "Ave." means average of three values. Superscripts a–b next to numbers indicate systems with which that value was statistically better ($p < .01$). **Bold** font represents top score for each evaluation criterion.

	No. of letters	No. of words
Vanilla	14.05	8.50
Fine-tuned	32.06	19.02
Gold	34.69	20.16

Table 6: Statistics of system utterances generated by models used for comparison. Gold means expanded system utterances (manually created).

Compared with manually created utterances (Gold), the naturalness score for the fine-tuned model was 0.71, and that of Gold was 0.56 (Table 3); this means that when considering naturalness alone, the fine-tuned model exceeded human-level performance, possibly because the model placed more emphasis on generating fluent utterances, rather than incorporating user information. Although the size of the SUI corpus is not very large, we found that the model fine-tuned with the SUI corpus successfully generated reasonable utterances. Note that the vanilla model generated natural and fluent utterances on the basis of dialogue context, but it did not have the capability to incorporate user information into utterances.

Table 6 shows the statistics of the system utterances generated by each model in the test set, and we found that the utterances generated by the fine-tuned model had almost the same length as the expanded system utterances (i.e., Gold). Table 7 shows examples of system utterances generated by each model. In the utterances generated by the fine-tuned model and Gold, "drama" from the dialogue context and "cake" or "sweets" from the user information were used in the generated content. This example shows that the fine-tuned model successfully generated natural utterances on the basis of user information and dialogue context.

To investigate whether the quality of generated utterances varies in accordance with the similarity between contexts and user information, we analyzed the relationship between user evalua-

Dialogue context	If you have any drama recommendations, please let me know. [SEP]Marigold.[SEP] What is the drama about? [SEP]It is a story about a man who sees the future when he puts his hand in contact with a person who has been wounded and tries to change it. [SEP] What do you like about this drama? [SEP]Because of the beautiful ending. (もし, おすすめのドラマがあったら教えてください。[SEP] マリーゴールド [SEP] そのドラマはどんな内容なんですか?[SEP] 手のひら 主人公の人が手のひらを なんかどうしたらいいんやろ 主人公が倒れてる人と手を合わせると未来が見えてしまってその未来を変えようと頑張るお話 [SEP] 面白そうですね。そのドラマのどういったところが好きなんですか?[SEP] すごいきれいな終わり方やったからです)
User information	Do you have any recommended sweets? [SEP] Cake. (その中でもおすすめのスイーツはありますか?[SEP] はなんかケーキ)
Vanilla	I see. I'll watch it sometime. (なるほど。今度見てみますね。)
Fine-tuned	I am sure you watched that drama while eating your favorite cake. (お好きなケーキを食べながら、そのドラマをご覧になったのでしょうね。)
Gold	Do you watch dramas where the theme is sweets or gourmet food? (スイーツとかグルメがテーマのドラマはご覧になりませんか?)

Table 7: Example system utterances generated by models used for comparison. All utterances were originally in Japanese. English translations were done by authors.

Cosine similarity	$[-0.4, -0.2)$	$[-0.2, 0.0)$	$[0.0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$
Naturalness	0.00	0.81	0.71	0.72	0.63	1.00

Table 8: Statistics of average naturalness evaluation scores and cosine similarity scores between contexts and user information.

tion scores and similarity scores between contexts and user information. We extracted the sentence embeddings of contexts and user information using Sentence-BERT (Reimers and Gurevych, 2019) (we used `stsb-xlm-r-multilingual`) and then calculated the cosine similarity between their embeddings. Table 8 shows the relationship between the average naturalness evaluation scores of the fine-tuned model given by three workers and cosine similarity scores between contexts and user information. As a result, we found that the evaluation scores had little relevance to the similarity scores between contexts and user information. This is a good indication that the quality of generated utterances does not depend on the similarity between contexts and user information. As opposed to the work by Xu et al. (2022b) that incorporates user information only when the topics are similar, this confirms that our models can incorporate arbitrary user information into system utterances regardless of the degree of similarity between contexts and user information.

5. Live Interactive Experiment

Despite the positive results in the previous section, it was still not clear if our fine-tuned model would work effectively in an interactive dialogue system with users. Therefore, in this section, we developed a chat-oriented dialogue system incorporating our fine-tuned model and evaluated its effectiveness through a live interactive evaluation.

5.1. Systems for Comparison

We developed three chat-oriented dialogue systems; two of them were baselines, and one was a system based on our fine-tuned model.

Vanilla This model does not utilize user information to generate system utterances. Utterances are generated by the vanilla Japanese Transformer encoder-decoder dialogue model (Sugiyama et al., 2023). Although we used a model fine-tuned with Japanese PersonaChat in Section 4.1, here we used a model fine-tuned with Japanese EmpatheticDialogues because this leads to more coherent dialogue.

UInfoRule This model is a replication of the work by Tsunomori et al. (2019). Utterances are randomly generated by hand-crafted rules with a probability of 30% and by Vanilla with a probability of 70%. These ratios were determined in line with (Tsunomori et al., 2019). The rules use hand-crafted templates such as "By the way, you talked about [word], didn't you? Let's talk more about it." to be filled in with a word from user information. For example, given the user information "I go to concerts," the generated utterance would be "By the way, you talked about concerts, didn't you? Let's talk more about it." We manually selected words to be used from user information for the experiment.

UInfoGen (ours) Utterances are randomly generated by our fine-tuned model in Section 4 with a probability of 30% and by Vanilla with a probability of 70%. UInfoGen generates utterances regardless of the degree of similarity between dialogue

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
a. Vanilla	**5.46**	4.66	**5.48**^{bb}	3.66	3.52	**5.00**^{bb}	**5.12**^{bb}	**4.96**^{bb}
b. UInfoRule	4.96	3.92	4.64	**5.04**^{aa}	4.08	3.76	3.74	3.66
c. UInfoGen	5.28	**4.78**^{bb}	5.36^{b}	4.96^{aa}	**4.80**^{aa,b}	4.84^{bb}	4.72^{bb}	4.70^{bb}

Table 9: Results of interaction evaluation (7 is highest). Superscripts a–c next to numbers indicate systems with which that value was statistically better. Double letters (e.g., aa) mean $p < .01$; otherwise, $p < .05$.

contexts and user information. When the similarity is high, the behavior of UInfoGen becomes similar to that of (Xu et al., 2022b). Whether to use our fine-tuned model or Vanilla is decided randomly because (a) we wanted to apply our proposed method at any time during the conversation, (b) we wanted to use the same settings as the baseline rule-based system (Tsunomori et al., 2019), and (c) the optimal strategy to decide the timing for using user information has not been established.

5.2. Evaluation Settings

We used CrowdWorks to recruit 50 workers who conducted dialogues in a text-chat interface with the three systems. The order of the systems was randomized. The workers were instructed to read a dialogue displayed in the chat interface as their own past dialogue with the system. The dialogues covered five items of user information selected randomly from the SUI corpus (test set in Section 4.1). Then, the workers conducted a dialogue with each system lasting 15 turns (30 utterances in total). After each dialogue session, they evaluated the system by indicating their degree of agreement with the following questions using a seven-point Likert scale. The questions were modified versions of those used in (Tsunomori et al., 2019).[7]

- Q1: The utterances of this dialogue system are easy to understand.
- Q2: The utterances of this dialogue system are interesting and informative.
- Q3: This dialogue system sounds familiar.
- Q4: This dialogue system remembers the contents of the past dialogue.
- Q5: This dialogue system appropriately uses the contents of the past dialogue.
- Q6: The utterances of this dialogue system are natural.
- Q7: I want to talk to this dialogue system again.
- Q8: I am satisfied with this dialogue.

[7]We mainly added questions concerning the system's ability to remember and use user information.

5.3. Results and Analysis

Table 9 lists the results of the interactive evaluation. We used the Steel-Dwass multiple comparison test (Dwass, 1960) for the statistical test. UInfoGen was high overall, Vanilla was high except for Q4 (remembering) and Q5 (using past dialogue), and UInfoRule was high for Q4.

When we compare UInfoGen and UInfoRule, both had high scores for Q4 (remembering). This indicates that UInfoRule and UInfoGen made users feel remembered by the system. However, UInfoRule did not appropriately use past dialogue because it had a lower score for Q5 (using past dialogues). In addition, it had significantly lower scores for all questionnaire items except for Q1 (understanding) and Q4. We found that incorporating user information into utterances using simple templates without considering dialogue context lowers the overall score.

When we compare UInfoGen and Vanilla, both had equally high scores, and there was no significant difference between them for all questionnaire items except for Q4 (remembering) and Q5 (using past dialogue). For Q4 and Q5, UInfoGen was significantly better. This indicates that our model, fine-tuned by the SUI corpus, enabled a dialogue system to remember and utilize user information; our model worked effectively in an interactive dialogue system with users. In addition, UInfoGen was better for Q2 (informative). By utilizing user information, UInfoGen succeeded in incorporating more information into utterances.

For system utterances generated by UInfoGen using user information, we calculated the cosine similarity between the dialogue context (up to the last three turns) and each piece of user information in the same manner as Section 4.3. The average similarity score was 0.29. The cosine similarity threshold in the work by (Xu et al., 2022b) was set to 0.7. However, in this experiment, the percentage of similarity score above 0.7 was just 2%, and that above 0.5 was 10%, indicating the importance of being able to incorporate arbitrary user information in system utterances. Note that, although Xu et al. (2022b) used ERNIE (Sun et al., 2020) embeddings instead of Sentence-BERT to calculate cosine similarity, the range of similarity values should fall in a similar range.

6. Conclusion and Future Work

In this paper, to build a good relationship between systems and users by remembering and utilizing arbitrary user information naturally and actively, we constructed a novel corpus, the **S**ystem utterance based on **U**ser **I**nformation corpus (**SUI corpus**). This corpus takes into account both user information on various topics and dialogue context. We fine-tuned a model to generate system utterances using the SUI corpus and conducted a subjective evaluation. The results showed that our fine-tuned model could incorporate arbitrary user information into system utterances regardless of the current dialogue topic while retaining appropriateness for the context. In addition, we found that our fine-tuned model was effective in a live interactive dialogue system.

There is still much room for improvement, especially in incorporating our fine-tuned model into dialogue systems. Our model sometimes generates unnatural utterances to incorporate arbitrary user information. We want to analyze the timing for effectively incorporating user information in dialogues since we only had random choice, which is obviously not the optimal solution. In fact, there were instances where the generated utterances disrupted the conversation flow. We would like to explore methods for reducing transitions that clearly cause problems, aiming to mitigate their impact. We would also like to enable the automatic extraction of user information and evaluate the system in real-world settings.

We also want to test the application of large language models (LLMs). Recently, few-shot learning using pre-trained LLMs has been applied successfully in generating natural sentences while taking into account the given information (Kasahara et al., 2022; Lee et al., 2022; Liu et al., 2022; Han et al., 2022). Thus, we believe that LLMs applied with the SUI corpus by few-shot learning methods could generate more natural utterances while incorporating user information.

7. Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19H05692. We used the computational resources of the supercomputer “Flow” at the Information Technology Center, Nagoya University.

8. Ethical Considerations

All evaluations were approved by the research ethics committee of our institution. We employed workers using a crowdsourcing service in evaluations. We made sure that the workers were paid above the minimum wage.

9. Bibliographical References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, arXiv:2001.09977v3.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the global WordNet conference (GWC)*, pages 56–63.

Meyer Dwass. 1960. Some k-sample rank-order tests. *Contributions to probability and statistics*, pages 198–202.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, pages 13665–13676.

Jeffrey A. Hall. 2019. How many hours does it take to make a friend? *Journal of Social and Personal Relationships*, 36(4):1278–1296.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5114–5132.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task

description, datasets, and evaluation metrics. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3146–3150.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015. Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2243–2248.

Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2022. Building a personalized dialogue system with prompt-tuning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) Student Research Workshop*, pages 96–105.

Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. 2019. Multimodal dialogue data collection and analysis of annotation disagreement. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages 201–213.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics (ACL)*, pages 1317–1337.

Bilyana Martinovsky and David Traum. 2006. The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pages 11–16.

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

Deborah Richards and Karla Bransky. 2014. ForgetMeNot: What and how users expect intelligent virtual agents to recall and forget personal conversational content. *International Journal of Human-Computer Studies*, 72(5):460–476.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *CoRR*, arXiv:2208.03188v3.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691.

Kensuke Sugo and Masafumi Hagiwara. 2014. A dialogue system with knowledge acquisition ability from user's utterance. *Journal of Japan Society of Kansei Engineering*, 13(4):519–526. (In Japanese).

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 34, pages 8968–8975.

Yuiko Tsunomori, Ryuichiro Higashinaka, Takeshi Yoshimura, and Yoshinori Isoda. 2019. Chat-oriented dialogue system that uses user information acquired through dialogue and its long-term evaluation. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages 227–238.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5180–5197.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! Open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics (ACL)*, pages 2639–2650.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 270–278.