Design guidelines for developing systems for dialogue system competitions

Kazunori Komatani, Ryu Takeda, Keisuke Nakashima, and Mikio Nakano

Abstract Because dialogue system development involves a variety of factors and requires multifaceted consideration, design guidelines for such development would be helpful. Although a neural-based approach can be used, it requires a vast amount of dialogue data, which would take too much effort to collect in the case of a system for a specific, fixed-length dialogue. Furthermore, the system design should explicitly consider errors in automatic speech recognition and language understanding, because they degrade the user impression and are inevitable when the system talks with general users. Accordingly, we propose design guidelines for developing such dialogue systems. Systems developed with the aid of these guidelines took first place in two dialogue system competitions: the situation track of the second Dialogue System Live Competition and a pre-preliminary contest of the Dialogue Robot Competition. Our proposed design guidelines are to (1) make the system take initiative, (2) prevent dialogue flows from relying too much on user utterances, and (3) include in utterances that the system understands what the user said. We describe details and examples for the systems designed for each of the two competitions.

1 Introduction

Dialogue system research has enabled the development of not only task-oriented systems but also non-task-oriented ones. Many studies have applied an end-to-end neural network approach to develop an open-domain, non-task-oriented dialogue system [1, 18, 12, 20]. This approach is used to generate appropriate responses to

---

Kazunori Komatani, Ryu Takeda, Keisuke Nakashima
SANKEN, Osaka University, e-mail: komatani@sanken.osaka-u.ac.jp, rtakeda@sanken.osaka-u.ac.jp, nakashima@ei.sanken.osaka-u.ac.jp

Mikio Nakano
Honda Research Institute Japan Co., Ltd. (Currently with C4A Research Institute, Inc.) e-mail: mikio.nakano@c4a.jp

user utterances, including their contexts, which often results in user-initiative dialogues. That is, the approach mainly focuses on how correctly the system responds to user inputs. However, it requires a vast amount of dialogue data, which involves too much effort to collect in the case of developing a system for a specific, fixed-length dialogue. Another consideration is how to design the entire dialogue to give better user experiences, including the dialogue flows and expression of system utterances, by regarding the system not as "a machine that responds reflexively" but as a partner in joint action [4].

Accordingly, we propose design guidelines for developing dialogue systems for a specific, fixed-length dialogue. Systems that we developed with the aid of these guidelines won first place in two dialogue system competitions. In those competitions, the systems had to conduct dialogues with various users and give good impressions. Specific dialogue designs were needed because the situations were different from one in which the system responds passively and keeps the dialogue going as long as possible [9]. Specifically, in these situations, the system needed to naturally guide the user's utterances while continuing the dialogue and showing that the system understood what the user had said, rather than accepting any user utterances and correctly responding to them as a user-initiative dialogue. On the other hand, a naive design of system-initiative dialogues would lead to rigid dialogues and not give a good user impression. Therefore, our design guidelines are intended to provide practical insights into dialogue system development with similar goals, as well as development of neural-based end-to-end dialogue modeling.

Several guidelines have recently been discussed in the context of the user interface [13, 21]. Those studies discussed how a completed dialogue system should behave from the user viewpoint on the basis of Nielsen's heuristics [16]. Our proposed guidelines, on the other hand, are useful during system development.

In the dialogue system community, there have been many discussions of system design. Many of them start from the principles of conversation between human interlocutors known as Grice's cooperative principle [6]. The principles were extended for task-oriented and human-machine dialogues by considering the distinction between generic and specific principles [2]. Concrete interaction guidelines based on the principles were also shown [15], and more comprehensive design guidelines were published for voice user interfaces (VUIs) [17, 5]. All of these guidelines are mainly for task-oriented dialogues, in which almost all user utterances need to be correctly understood. In Contrast, our task is a little different: the system needs to establish a dialogue for a certain period of time while giving a good impression to the user.

Other strategies for increasing user initiative were recently proposed for an Alexa Prize bot and tested experimentally [7]. The authors of that study preferred longer user utterances because they assumed such utterances would facilitate more engaging conversations in their task. Here, the dialogue that we want to achieve is different: the tasks of our systems have specific goals, and the dialogues need to feel natural to the user.

The rest of this paper is organized as follows. Section 2 gives overviews of the two dialogue system competitions that we participated in. Section 3 describes the

Table 1 Overview of the two dialogue system competitions

	DSLC2 situation track	DRC
Task	Chat in designated situation	Tourist information
Length	15 utterances	5 minutes
Input modalities	Text	Speech and vision (optional)
Output modalities	Text	Speech and robot motion
Evaluation criteria	Humanness (appropriate to the situation)	Seven items listed in Table 2, including “Naturalness of dialogue”, “Satisfaction with dialogue”, “Quality of service”, etc.

proposed guidelines that aided us in developing the two systems for the competitions. More details on the individual systems are given in Sections 4 and 5, along with examples. Section 6 concludes the paper.

2 Dialogue System Competitions

We first give an overview of the two dialogue system competitions that we participated in:

- Situation track in the second Dialogue System Live Competition (DSLC2 situation track)[1]
- Dialogue Robot Competition (DRC)

Our team’s systems won first place in each competition. In both cases, the target language was Japanese.

Table 1 summarizes the main characteristics of the two competitions. The dialogues were conducted under specific situations and had fixed lengths. The dialogues in DSLC were text chats with the dialogue systems, while those in DRC were spoken dialogues with an android robot.

2.1 DSLC2 situation track

The DSLC is a competition in which an audience watches and evaluates live dialogues between users and dialogue systems [8]. The dialogues are conducted as online text chats on Telegram[2]. A screenshot is shown in Fig. 1. After preliminary selection via crowdsourcing, three systems proceeded to the live event, performed dialogues with designated users who had been selected by the organizer, and were evaluated by the audience. The second edition took place in autumn 2019 and had two tracks: the open track, and the situation track, in which our team participated.

[1] https://dialog-system-live-competition.github.io/dslc2/ (written in Japanese)
[2] https://telegram.org/

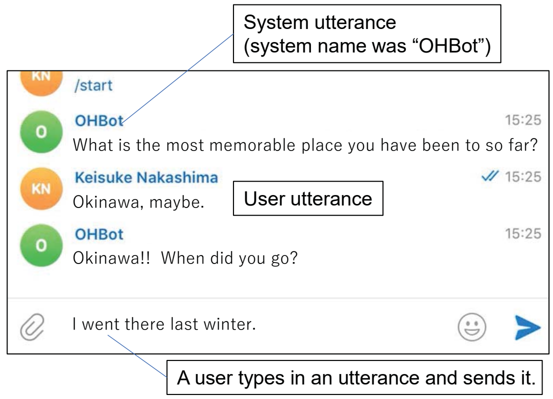

Fig. 1 Screenshot of the DSLC system (translated from Japanese)

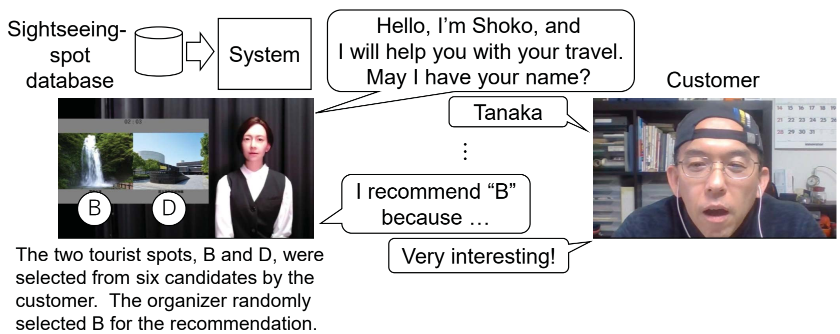

Fig. 2 Overview of DRC task

The situation track used the following setup: "The user and the system are friends from their school days, and they start chatting on topics related to the most memorable trips and places they have been." System developers were allowed to specify their system's gender; that is, they could select male-to-male or female-to-female dialogues. We selected female-to-female dialogues. The length of a dialogue was 15 exchanges (i.e., pairs of user and system utterances). The track's evaluation criterion was "how human (appropriate to the situation) was the conversation the system conducted."

2.2 DRC

The DRC was held to promote improvement in the spoken dialogue technologies of android robots. In the DRC's dialogues, the robot acted as a travel agent and recommended a tourist spot to recruited participants acting as customers. The dialogues thus required the robot to provide information on tourist spots and hospitality that would satisfy the customer. Note that the customer interacted with the robot by voice, while the robot could speak and move its hands and head during a dialogue.

An overview of the DRC task is shown in Fig. 2. First, a participant acting as a customer selected two tourist spots from six candidates in advance. The competition organizer randomly specified one of the two spots that the robot should recommend. During the dialogues, the robot sought to persuade the participant to be interested in the specified spot. Pictures of the two spots were shown. The dialogue duration was five minutes. Each system was evaluated through questionnaires submitted by the participants after the dialogues.

The organizer provided basic modules for input and output to conduct the dialogues, such as speech-to-text, text-to-speech, and robot motion control modules. This enabled system developers to focus on the core dialogue design, while they could use their own recognition modules if they wanted. The knowledge of tourist spots was also provided as a database in advance.

3 Proposed Design Guidelines

As listed in Table 1 above, the competitions had the following important characteristics:

- The dialogues were of fixed length: 15 turns for DSLC2 and five minutes for DRC.
- The evaluation criteria explicitly included the user's impression.

Therefore, we needed a reasonable design that could give the user a good impression while establishing the dialogue in a particular situation, instead of an approach of collecting a huge amount of data and training the system to accept all kinds of user utterances. In particular, erroneous system utterances in such short dialogues would have fatally degraded the user's impression.

Accordingly, we developed three key guidelines from various aspects of the dialogue design requirements, as follows:

1. Make the system take initiative
2. Prevent dialogue flows from relying too much on user utterances
3. Include in utterances that the system understands what the user said

3.1 Make the system take initiative

The system should avoid being questioned by the user as much as possible. The reason is that the system is obligated to respond when the user asks a question, but it is very difficult to respond appropriately to all types of questions. Specifically, we sought to end the system's turn by asking a question or making an utterance that would elicit empathy from the user[3].

A well-known finding from research on task-oriented dialogue systems is that novice users prefer system-initiative dialogues to user-initiative dialogues [11]. That finding also supports this guideline, because a first-time user can easily proceed a dialogue by following the system's utterances, especially when the user does not know what the system can do. In addition, by ending an utterance in the form of a question, the system can explicitly indicate that it is the user's turn to speak, making it less likely that the user becomes confused about taking turns with the system.

At the same time, the system needs to avoid asking too many questions or speaking at length without ending its turn. Therefore, we divide the entire dialogue into several phases, which prevents the dialogue from becoming monotonous and enables us to develop each phase independently.

3.2 Prevent dialogue flows from relying too much on user utterances

We design dialogue flows to not rely too heavily on user utterances when possible. This can be the case when the subsequent dialogue can be established regardless of the user's response, and it is effective when the user's response is difficult to predict because the system's question has many possible answers (e.g., an open-ended question). It is impossible for a dialogue system to understand every possible user utterance correctly, although such systems are often expected to have this capacity. Moreover, even if a system could understand every possible user utterance correctly, it would be difficult to establish a system response for every possible user utterance. For this reason, an approach based on end-to-end machine learning has recently been preferred, but, in this approach, the dialogue flow is left to machine learning results trained with a vast amount of data. It seems impossible to perfectly solve this problem and control what the system says. Confidence measures have been used to reject candidates with low confidence [17], but errors cannot be completely avoided even in that case.

Therefore, we design system responses to be valid even when the system does not correctly understand a user utterance[4]. Figure 3 shows an example of such a system utterance (S15). In both cases, the system asks a question (S14), but the user responses (U14) differ: the user is interest in the system's offer in the upper example but not in the lower example. The system response (S15) seems valid in either case.

[3] A similar guideline, "Avoid system utterances that may induce user questions," was also listed as a design guideline in our previous framework for developing closed-domain chat dialogue systems [14].

[4] This approach is from a lecture given by Dr. Iio before DSLC2. The video (in Japanese) is available at https://dialog-system-live-competition.github.io/dslc2/lecture.html. It is part of know-how shared in Prof. Ishiguro's Laboratory at Osaka University, where he previously worked and where several talking robots were developed.

S14:	How about going there together?
U14:	Great. When are we going?
S15:	Let's talk over dinner together sometime, then.
S14:	How about going there together?
U14:	I guess I'm too busy for that.
S15:	Let's talk over dinner together sometime, then.

Fig. 3 Example of a system utterance S15 that does not depend on the preceding user utterance U14. The upper and lower examples are different dialogues. In this and subsequent examples, labels starting with S or U denote system or user utterances, respectively.

For the competitions, the dialogue flows were designed by hand. To give a good impression in a fixed-length dialogue with designated tasks, this is a more reasonable approach than collecting a huge amount of dialogue data for each task and training a neural model to obtain such dialogue flows.

3.3 Include in utterances that the system understands what the user said

Adding language understanding (LU) results of a user utterance to the system utterance tends to result in a good user impression [10]. The guideline discussed in Section 3.2 corresponds to ignoring user utterances, but if the system completely ignored the user utterances, it would degrade user impression.

Therefore, we include the LU result of the user utterance in the system utterance when the system seems to correctly understand it. This often becomes possible when the system asks a more specific question, rather than a vague one, and the LU result is quite likely to be correct, e.g., it matches the expected entries in the dictionary. Although this guideline conflicts with the second guideline, we aim to apply it when possible so that the user will feel that the system is taking into account what he or she has said, thus giving a better impression.

In addition to simply adding the LU result of a user utterance (e.g., the user's name, the transportation method to get there, etc.) as it is, we also use domain ontology of place names to show understanding. Figure 4 shows this in a dialogue example. For DSLC2, we prepared an ontology of famous sightseeing spots including Ishigaki Island, which is part of Okinawa prefecture. In the example, this enables the system to express its understanding via "You've been to Okinawa!" in S2, which is better than just repeating a literal word in the user utterance.

Implicit confirmation requests in task-oriented dialogues have also been known to include the LU result of a user's utterance in the next system utterance [19]. The goal here is not to be certain of the correctness of the LU result, but to give the user a good impression, i.e., to show that the system understands what the user said. The tourist information task can be basically performed even if the system only speaks about tourist information in one way, but remembering what the user said can be

S1:	What is the most memorable place you have been to so far?
U1:	Ishigaki Island, maybe.
S2:	You've been to Okinawa! The beaches are so beautiful and amazing! When did you visit there?

Fig. 4 Example of a system utterance evoking understanding of place names (Ishigaki Island is part of Okinawa)

considered to show that the system is treating the user well; this is an aspect of hospitality and thus improves the user impression.

4 System Design and Results for DSLC2 Situation Track

4.1 Specific designs

4.1.1 Phase design

We prepared three phases for this task of 15 utterances, i.e., about five utterances for each phase. The task was a chat with a female friend about memorable trips and places.

In the first phase, the system asked the user in turn about several "slots," such as the place of her most memorable trip, when she visited, her impression, and what she ate there. The system skipped asking about a slot if it had already been mentioned. For example, the question "When did you visit there?" was skipped if the user had already mentioned when she visited a particular place, e.g., "I went to Okinawa last summer." This was judged on the basis of the LU result of the user utterance.

In the second phase, the system talked about its experience on its own trip. This phase kept the dialogue from becoming monotonous and prevented the system from continuing to ask questions. We also inserted a quiz in this phase to stimulate the user's interest and prevent the system from continuing to speak without ending its turn.

In the final phase, the system discussed future plans with the user. The system talked about a place where the user wanted to go and also invited the user to go together to a place where the system was planning to travel.

We also made the system avoid talking about the same place in the first and second phases. This was to prevent the same content from being repeated, which could make the user suspicious that the system does not understand the dialogue. By default, the system talked about its experience when it went to Okinawa in the second phase, but the topic changed to Hokkaido (another popular place) if the user had already talked about Okinawa in the first phase. This was judged on the basis of the LU results and the system's ontology (i.e., whether the place name in the first phase was part of Okinawa). Figure 5 shows an example of such avoidance. In the first phase (U1), the user talked about Miyako Island, which the system recognized

S1:	What is the most memorable place you have been to so far?
U1:	Miyako Island was great.
...	...
S6:	I see. As for me, I've recently become addicted to sweets from Hokkaido :D

Fig. 5 Example of the system avoiding the same topic in the first and second phases. Emoticons are denoted by ":D".

S1:	What is the most memorable place you have been to so far?
U1:	Well, it's all too common, but maybe Hawaii.
S2:	Hawaii!! When did you go there?
...	...
S14:	... How about going there together?
U14:	I can't wait to go!
S15:	Let's talk over dinner together sometime, then :D

Fig. 6 Excerpt of a dialogue log for our system during the live event of DSLC2

as part of Okinawa prefecture by using its ontology. As a result, the system talked about Hokkaido in the second phase (S6).

4.1.2 Showing empathy and character consistency

We also added expressions to show empathy at the beginning of an utterance after asking the user a question and getting a response from her. These expressions included "right," "indeed," "I know." This was to prevent the system from being perceived as not listening to the user. The system changed the expressions to avoid degrading the user's impression by repeating the same one.

Because the task setting was a chat between close female friends, we took care to ensure that the expressions and content of utterances and the use of emoticons (denoted by ":D" in this paper) were consistent with how a Japanese woman would use them. The consistency was checked by a female developer and another woman several times.

4.2 Results and examples

Our system received the highest score among the seven systems that participated in the preliminary round of the situation track of DSLC2. It received an average score of 4.1 on a five-point Likert scale from 50 crowdworkers. The criterion was "how human (appropriate to the situation) was the conversation the system conducted."

After proceeding to the final live event of DSLC2, our system won first place through evaluation by the audience of about 100 people including dialogue system researchers. The logs are available to the public and include the evaluated label

S13:	I'm actually thinking of going to Shinshu next spring.
U13:	Shinshu is also nice. What are you doing there, skiing or snowboarding?
S14:	There are lots of yummy things, like apple sweets :D My friend's parents' house is ...

Fig. 7 Example of a failure due to incorrect type estimation

distributions for each utterance in the live event[5]. The best distributions were given to the two system utterances S2 and S15, shown in Fig. 6: 91.6% of the audience gave the highest label of "Good", on a three-point scale. Utterance S2 showed the LU result of the place name "Hawaii," as mentioned in Section 3.3. Utterance S15 was similar with those in Fig. 3: the system utterance did not depend on the previous user utterance, as discussed in Section 3.2.

There were several utterances that seemed inappropriate in the context, however. Figure 7 shows an example of such a failure in the preliminary round. Utterance U13 was wrongly assigned the same dialogue-act type as the question "What is in Shinshu?" by the LU module. As a result, the system started the following utterance S14 with a prepared sentence about a local specialty. This kind of mistake could be avoided by adding a correct dialogue-act type to the LU module and training it, but such errors are inevitable.

5 System Design for DRC and Results

5.1 Specific designs

5.1.1 Phase design

A dialogue system for this task was supposed to provide information on tourist spots to a customer and take the customer's travel request. We divided the entire dialogue of five minutes into four phases: (1) introduction, (2) explanation of tourist spots, (3) recommendation, and (4) Q&A.

In the first phase, the system gave the customer greetings and simple questions as ice breakers. The questions were about 1) the customer's experience talking to robots, 2) the customer's name, 3) the transportation method to be used, and 4) any traveling companions. The system stored the customer's answers for use in a later phase. If automatic speech recognition (ASR) or LU failed, a default value was then used[6]. For example, if the system could not recognize the customer's name, it instead called the customer "Sir" or "Madam".

In the second phase, the system gave an outline of the two tourist spots that the customer had selected and then gave him or her more specific information of the two spots. Specifically, the system gave descriptions of the spots, explained how to

[5] https://dialog-system-live-competition.github.io/dslc2/result.html

[6] This is similar to the "MoveOn" strategy [3].

	(robot turns toward the pictures on display)
S:	The left picture shows Minoh Falls, and the right one shows Soji-ji temple.
	(robot turns toward the participant)
S:	Which picture, right or left, impresses you more?
U:	The right one.
S:	Oh, nice. I also think so.
...	...

Fig. 8 Question about two pictures of tourist spots in the explanation phase.

S:	Do you like spots in the "rivers, canyons, and water falls" category?
U:	Yes.
S:	Then I strongly recommend that you visit Minoh Falls. You can feel relaxed and comfortable.
S:	Also, you said "My impression of the picture is good."
...	...

Fig. 9 Examples of recommendation reasons in the recommendation phase

access them, and mentioned categories such as "temples and shrines," "factories and facilities," and so on. The system also asked the customer which picture he or she preferred between pictures of the two spots, as shown in Fig. 8. The answer to this question was used as a reason for the recommendation in the following phase.

In the third phase, the system recommended one of the two tourist spots and explained the reasons. We prepared sentences with recommendation reasons in advance, and the system selected them according to what the customer said during the dialogue, such as his or her preference between the two pictures. The most specific reasons we had prepared were related to the customer's preference for the touring spot's category, which the system asked during the dialogue, as shown in Fig. 9. Giving more specific reasons would be better because a customer would not be convinced by general reasons that are not related to specific spots; however giving more detailed reasons would require complicated ASR and LU technologies to understand the customer's preferences and experiences, which would increase the risk of misunderstanding.

In the final "Q&A" phase, the system answered questions from the customer as long as time permitted. Because it was difficult to answer the customer's open-ended questions, the system gave the customer several examples of what kinds of questions could be answered. When five minutes had passed from the beginning of the dialogue, the system ended dialogue with closing remarks to the customer.

5.1.2 Strategies in LU and turn taking

Our strategy to reduce misunderstandings caused by LU failures was to show a few words or phrases as examples in each system question. This was because our LU approach was based on pattern matching between recognized character sequences and a prepared word set. We expected customers to utter one of the examples as the answer. For example, when the system asked "Are you planning to use a private car, train, or other means of transportation?”, the customer’s possible answers were almost entirely restricted to “private car,” “train,” and several other words. We also added similar and possibly misrecognized expressions to the original set of expected words and phrases.

We also manually designed the timing of when the system accepted a customer utterance, i.e., whether the customer was allowed to barge into system utterances. For example, to avoid unexpected situations, the system basically did not accept any customer utterances while it was explaining something. In other words, the system only attempted to understand customer responses to its explicit questions.

An elapsed time after each system utterance was used to maintain turn taking even when ASR or LU failed. After a certain amount of time had elapsed, the system was designed to say something. Without this capability, if the system could not detect a customer’s utterance while waiting for an answer, both the system and the customer might have had to wait and, the silence would have continued; instead, the system made a confirmation utterance about the current situation or moved on to the next utterance.

5.1.3 Speech synthesis and robot motions

The pronunciations of the system utterances and the speaking speed were carefully checked in advance. Unlike in text chats, these factors are important because they affect the customer impression. For example, if the speaking speed is extremely fast or slow, customers may feel stressed.

Coordination of the robot motion with the utterance is also important, because it would be strange if the robot did not move at all. Accordingly, the robot shook its body slightly and slowly, and it blinked its eyes by default. We also created two specific motions: bowing upon greeting the customer and turning toward the display when showing the pictures of the two touring spots. The latter motion was designed to create joint attention by guiding the customer’s eyes to the pictures. A dialogue example with this kind of motion is shown in Fig. 8.

5.2 Results of pre-preliminary contest and examples

The DRC’s pre-preliminary contest was held in March 2021. Nine systems were evaluated, including the organizer’s baseline system. Because of the COVID-19 pandemic, the recruited participants performed dialogues with the android robot via remote software. Each participant had a maximum of one to three dialogues and filled out a questionnaire after each dialogue. Each system was scored by about 10 participants.

The questionnaire items were prepared by the organizer and are listed in Table 2. Each item was scored on a seven-point scale. The table also lists the average scores

Table 2 Questionnaire items and average scores of the systems.

Questionnaire items	Our system	Baseline
1. Satisfaction with recommendation	5.9	4.5
2. Amount of provided information	5.7	4.9
3. Naturalness of dialogue	4.9	4.0
4. Appropriateness	5.7	4.2
5. Satisfaction with dialogue	5.3	4.0
6. Quality of service	6.0	4.4
7. Usefulness of provided information	6.3	4.6

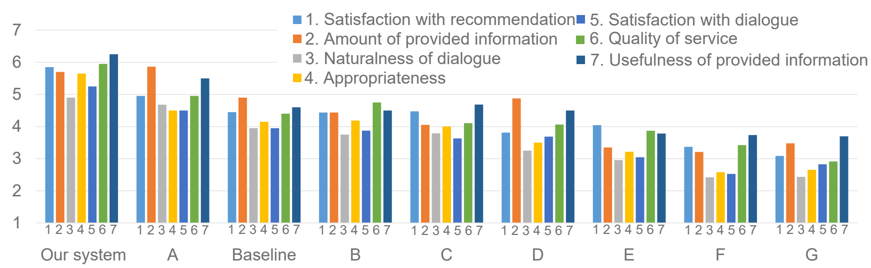

Fig. 10 DRC Results. The vertical axis represents the average user impression on a seven-point scale, with a higher score indicating a better impression.

of our system and the baseline for each item. The score gaps between them were typically over one point.

Figure 10 shows the score distributions for all items and systems. The letters A to G represent the systems developed by the other teams. Our system received the highest scores for most of the questionnaire items, while the baseline was ranked third. Note that these scores were not relative among the systems because each system was evaluated by different participants.

The better-scored questionnaire items for our system were related to providing information on the tourist spots and the appropriateness of the system response: (1) satisfaction with recommendation, (2) amount of provided information, (4) appropriateness, (6) quality of service, and (7) usefulness of provided information. The scores of these items were better because our system had a specific phase for explaining the tourist spots. The system's responses were considered appropriate because the system was designed and tuned manually to avoid unexpected system behaviors. It was thus natural that these scores were relatively high, because we manually prepared guidance for the system on the task of providing tourist information. In addition, the robot's motions of bowing, nodding, and turning toward the display to create the joint attention might have resulted in better scores for the appropriateness of the system's behavior.

The relatively worse-scored items were related to the naturalness of the dialogues: (3) naturalness of dialogue and (5) satisfaction with dialogue. We suggest that this was partly because our system did not allow the customer to speak freely during system utterances. This was an inevitable choice to avoid dialogue failures,

time		
239.0	S31:	I want you to go to both places, but I recommend Meiji-Naruhodo Factory, Osaka.
241.7	U10:	Thanks.
249.3	S32:	First, you can enjoy sightseeing and talking with your friends.
251.3	U11:	Why? # This was rejected and ignored because it was during system utterance.
255.9	S33:	Do you like factory and facility tours?
264.8	U12:	Yes, I like them.

Fig. 11 Example of a failure due to the system not allowing the customer to barge in during system utterance

time		
37.7	S7:	First, may I have your name?
39.1	U5:	Yes.
44.0	U6:	Uh, I'm XXX. (*E, XXX desu*)
46.7	S8:	Your name is XXX, right?
52.5	U7:	*Yoi desu*
53.8	S9:	Your name is Yoi, right?
		# "*Yoi*" means "right," but it was not in the dictionary and was misunderstood as a name.
59.6	U8:	No, I'm XXX. (*Ie, XXX desu*) # This ASR result was obtained at 62.3.
61.7	S10:	I'm sorry, but let me skip confirming your name because it was hard for me to hear.
68.0	S11:	Sir, which are you going to use, a private car or public transportation, during your travel?

Fig. 12 Example of a failure of name recognition. XXX was the customer's name but is anonymized here for privacy. Text in *italics* represents literal Japanese transcriptions.

given the current performance of ASR and LU technologies: it would have become more difficult to correctly understand the customer's utterances and continue the dialogue if the system allowed the customer to barge in and tried to recognize every utterance. Figure 11 shows an example. Here, all ASR results were rejected while the system was speaking; that is, the question from the participant (U11) was ignored. This behavior might have worsened the score for the naturalness of dialogue.

In addition, Fig. 12 shows an example in which the name recognition failed. The participant replied "That's right (*Yoi desu*)" as U7, but the system misunderstood "*Yoi*" as the participant's name because it was not included in the dictionary as a variant expression of "right (*yoi*)" for LU based on pattern matching. Furthermore, the user utterance U8 was ignored because barge-in was not allowed: its ASR result was obtained at a time of 62.3, which was just after the system had started speaking S10 (at 61.7). Such behavior might also degrade the naturalness of dialogue.

6 Discussion and conclusion

We have proposed design guidelines for developing dialogue systems for competitions. Our systems developed with the aid of these guidelines won first place in two competitions.

The guidelines here correspond to a previous experimental result from a user impression analysis of a chat dialogue system in the food and restaurant domain [14]. That study showed correlations between the main questionnaire item ("I'm willing to chat with the system again") and seven other items. The three items with higher correlations were "The dialogue was fun," "The dialogue was natural," and "The system understood my utterances."[7] The second item corresponds to one of our design guidelines, namely, the guideline to prevent dialogue flows from relying too much on user utterances (Section 3.2) in order to avoid disruptions by ASR and LU errors and make the dialogue as natural as possible. The third item corresponds to our guideline to include in utterances that the system understands what the user said (Section 3.3). On the other hand, the first item may depend heavily on the dialogue content, which is beyond the scope of this paper.

It would be good to quantify the impact of the proposed design guidelines, but that would require a new experimental design, and it is thus beyond the scope of this paper. Nevertheless, we hope that these design guidelines will inspire developers of other dialogue systems.

Acknowledgements This work was partly supported by JSPS KAKENHI Grant Numbers JP19H05692 and JP19H04171, and JST, PRESTO Grant Number JPMJPR1857, Japan.

References

[1] Adiwardana D, Luong M, So DR, Hall J, Fiedel N, Thoppilan R, Yang Z, Kulshreshtha A, Nemade G, Lu Y, Le QV (2020) Towards a human-like open-domain chatbot. CoRR URL https://arxiv.org/abs/2001.09977, 2001.09977

[2] Bernsen NO, Dybkjaer H, Dybkjaer L (1996) Principles for the design of cooperative spoken human-machine dialogue. In: Proc. International Conference on Spoken Language Processing (ICSLP), vol 2, pp 729–732 vol.2, DOI 10.1109/ICSLP.1996.607465

[3] Bohus D, Rudnicky AI (2005) Sorry, I didn't catch that! - an investigation of non-understanding errors and recovery strategies. In: Proc. SIGdial Workshop on Discourse and Dialogue, pp 128–143

[4] Clark HH (1996) Using Language. Cambridge University Press

[5] Deibel D, Evanhoe R (2021) Conversations with Things: UX Design for Chat and Voice. Conversations with Things: UX Design for Chat and Voice, Rosenfeld Media, URL https://books.google.co.jp/books?id=BlouzgEACAAJ

[6] Grice HP (1975) Logic and conversation. In: Cole P, Morgan JL (eds) Syntax and Semantics: Vol. 3: Speech Acts, Academic Press, New York, pp 41–58, URL http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf

[7] The remaining four items with lower correlations were as follows: "The dialogue went well," "The system was polite," "The system was friendly," and "The system did not often change the topic." These items were of relatively lower importance, given the current system performance.

[7] Hardy A, Paranjape A, Manning C (2021) Effective social chatbot strategies for increasing user initiative. In: Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp 99–110, URL https://aclanthology.org/2021.sigdial-1.11

[8] Higashinaka R, Funakoshi K, Inaba M, Tsunomori Y, Takahashi T, Akama R (2019) Dialogue system live competition: Identifying problems with dialogue systems through live event. In: Marchi E, Siniscalchi SM, Cumani S, Salerno VM, Li H (eds) Proc. International Workshop on Spoken Dialogue System Technology (IWSDS), Springer, Lecture Notes in Electrical Engineering, vol 714, pp 185–199, URL https://doi.org/10.1007/978-981-15-9323-9_16

[9] Hu S, Liu Y, Gottardi A, Hedayatnia B, Khatri A, Chadha A, Chen Q, Rajan P, Binici A, Somani V, Lu Y, Dwivedi P, Hu L, Shi H, Sahai S, Eric M, Gopalakrishnan K, Kim S, Gella S, Papangelis A, Lange P, Jin D, Chartier N, Namazifar M, Padmakumar A, Ghazarian S, Oraby S, Narayan-Chen A, Du Y, Stubell L, Stiff S, Bland K, Mandal A, Ghanadan R, Hakkani-Tur D (2021) Further advances in open domain dialog systems in the fourth alexa prize socialbot grand challenge. In: Alexa Prize Proceedings

[10] Kobori T, Nakano M, Nakamura T (2016) Small talk improves user impressions of interview dialogue systems. In: Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp 370–380, URL https://www.aclweb.org/anthology/W16-3646

[11] Komatani K, Ueno S, Kawahara T, Okuno HG (2005) User modeling in spoken dialogue systems to generate flexible guidance. User Modeling and User-Adapted Interaction 15(1):169–183

[12] Komeili M, Shuster K, Weston J (2021) Internet-augmented dialogue generation. CoRR URL https://arxiv.org/abs/2107.07566, 2107.07566

[13] Langevin R, Lordon RJ, Avrahami T, Cowan BR, Hirsch T, Hsieh G (2021) Heuristic evaluation of conversational agents. In: Proc. Conference on Human Factors in Computing Systems (CHI), URL https://doi.org/10.1145/3411764.3445312

[14] Nakano M, Komatani K (2020) A framework for building closed-domain chat dialogue systems. Knowledge-Based Systems 204:106212, DOI https://doi.org/10.1016/j.knosys.2020.106212

[15] Niculescu A (2011) Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality. PhD thesis, University of Twente, URL https://research.utwente.nl/en/publications/conversational-interfaces-for-task-oriented-spoken-dialogues-desi, sIKS Dissertation Series ; no. 2011-49

[16] Nielsen J, Molich R (1990) Heuristic evaluation of user interfaces. In: Proc. Conference on Human Factors in Computing Systems (CHI), p 249–256, DOI 10.1145/97243.97281, URL https://doi.org/10.1145/97243.97281

[17] Pearl C (2016) Designing Voice User Interfaces: Principles of Conversational Experiences, 1st edn. O'Reilly Media, Inc.

[18] Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, Xu J, Ott M, Smith EM, Boureau YL, Weston J (2021) Recipes for building an open-domain chat-

bot. In: Proc. European Chapter of the Association for Computational Linguistics (EACL), pp 300–325, URL https://aclanthology.org/2021.eacl-main.24
[19] Sturm J, den Os E, Boves L (1999) Issues in spoken dialogue systems: Experiences with the Dutch ARISE system. In: Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, Kloster Irsee, Germany, pp 1–4
[20] Xu J, Szlam A, Weston J (2021) Beyond goldfish memory: Long-term open-domain conversation. CoRR URL https://arxiv.org/abs/2107.07567, 2107.07567
[21] Yang X, Aurisicchio M (2021) Designing conversational agents: A self-determination theory approach. In: Proc. Conference on Human Factors in Computing Systems (CHI), URL https://doi.org/10.1145/3411764.3445445