Interspeech 2024
1-5 September 2024, Kos, Greece

Investigation of look-ahead techniques to improve response time in spoken dialogue system

Masaya Ohagi[1], *Tomoya Mizumoto*[1], *Katsumasa Yoshikawa*[1]

[1]SB Intuitions Corp

{masaya.ohagi, tomoya.mizumoto, katsumasa.yoshikawa}@sbintuitions.co.jp

Abstract

This paper reports a new method that improves the response speed in spoken dialogue systems that use large language models. In existing systems, the start of the chatbot's response after the user utterance is delayed by the time required to generate that response. In contrast, our system predicts what the user may say next and pre-generates the bot's response before the user finishes speaking. This look-ahead technique allows the response to be returned by simply matching the predicted user utterance with the actual user utterance. Evaluation results show that our method has high look-ahead accuracy in task-oriented dialogue, contributing to improved response speeds.

Index Terms: dialogue system, look-ahead

1. Introduction

With the development of large language models (LLMs) such as ChatGPT [1] and the high-performance Speech2Text [2] and Text2Speech [3] models, systems capable of fluent natural spoken dialogue are becoming increasingly feasible. Although spoken dialogue systems have a variety of applications, such as smart assistants and customer support, one drawback with LLMs is the time required to generate a response. For example, OpenAI's GPT-3.5-turbo can only generate 14 tokens per second [4]. In some languages, such as Japanese, speakers tend to begin responding to an utterance within one second [5]. Hence, the existing system which generates responses after the user utterance leaves an unnatural gap before the bot starts speaking, leading to turn-taking failures and user confusion.

An existing solution is to vocalize each token as it is generated, instead of vocalizing the response after the LLM has finished generating all the sentences. However, this approach is incompatible with the post-processing, such as the filtering of harmful utterances. If we vocalize tokens before the whole sentence has been generated, when the harmful content is identified, part of the sentence has already been vocalized. Thus, this research focuses on a spoken dialogue system that vocalizes utterances after the completion of utterance generation, but with an improved response speed. For this purpose, we propose a system that predicts the user's utterance and pre-generates the bot's response. We call this combination of prediction and pre-generation "look-ahead".

A schematic diagram of the proposed method is shown in Figure 1. In the existing system, for example, after the bot asked "What transportation do you use?", the bot simply waited for the user to respond "A car, I guess". We allocate this idle time to the look-ahead process which consists of predicting the user utterance and pre-generating a bot response to it. Our system predicts the user's utterance as "I'll go by car" or "I'll go by train" and pre-generates a response for each predicted utterance.

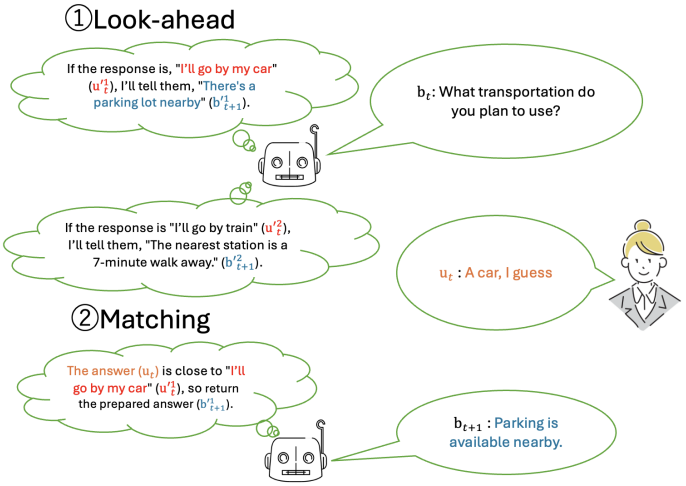

Figure 1: *Proposed system predicts the user's utterance and generates the bot's response to prediction in advance.*

This look-ahead allows the bot to have several responses ready when the user finishes speaking. After the user utterance, the predicted user utterances are compared with the actual user utterance. If the prediction succeeds, the pre-generated response "Parking is available nearby" will be returned only with the matching time between the predicted utterances and the actual one. Therefore, this approach can reduce the generation time after the user utterance compared with existing systems.

We evaluate our system using two types of real dialogue data between a bot and a human. The results show that our system predicts approximately 20% of the user's utterances with a high degree of similarity. In addition, the measured look-ahead times confirm that the look-ahead technique can be completed before the user finishes their speech. This result guarantees the effectiveness of our system in improving the response speed of spoken dialogue systems.

2. Related work

There are two main types of dialogue systems: rule-based [6] and generative model-based [7]. A larger generative model requires longer to generate responses, making the response time problematic [8]. We propose a look-ahead technique as a solution for this problem.

There are several existing studies on dialogue look-ahead techniques. Several achieve efficient task resolution by performing a look-ahead process [9, 10, 11], while others use look-ahead as supplementary information to resolve other tasks

10.21437/Interspeech.2024-67

```
### Instruction
Based on the ``conversation history'' between
  the user and the bot, predict 10 patterns of
  user utterances and bot utterances.

### Example: Dialogue History
Bot: Do you like to visit historical
  institutions?
### Example: Output
1: User: I like that. Bot: Ok, I will focus on
  temples, shrines, and other historical
  facilities.
2: User: No, I don't much like that. Bot: Ok,
  sir. Let me introduce you instead to some
  places where you can enjoy nature and museums.

### Dialogue History
Bot: I am a tourist guide, what do you do?
User: Editor.
Bot: Editor! That's great. What exactly do you
  do as an editor?
### Output
```

Figure 2: *Prompt for look-ahead module*

[12, 13]. However, there have been no attempts to introduce the look-ahead technique as a means of reducing the response generation period in real-time spoken dialogue systems. To deal with real-time spoken systems, we must tackle the restriction of finishing the look-ahead process before the end of the user's speech and the matching of actual and predicted user utterances. In this respect, the task addressed in this research has not been considered by previous studies.

3. Proposed method

Our system consists of two modules: 1) **a look-ahead module** that predicts the user's next utterance and pre-generates a response, and 2) **a matching module** that verifies how similar the predicted user utterance is to the actual user utterance and returns the pre-generated response if it is sufficiently similar. From here, we define a combination of a bot's utterance and a user's response as a "turn". The bot's utterance in turn t is denoted as b_t, and that of the user is u_t.

First, the look-ahead module predicts N patterns of user utterance u_t based on the dialogue history at dialogue turn t. Let $u_t^{'k}$ denote each predicted user utterance ($k = 1, 2, ..., N$). Also, for each predicted user utterance $u_t^{'k}$, we pre-generate the bot response $b_{t+1}^{'k}$. These prediction and pre-generation processes are expected to have been completed by the time the bot vocalizes utterance b_t and the user finishes speaking u_t. Therefore, if $u_t^{'k}$ correctly predicts user utterance u_t, utterance $b_{t+1}^{'k}$ is ready to be returned immediately as a response. This study used GPT-3.5-turbo to predict user utterances and simultaneously pre-generate bot responses. The prompt for the look-ahead module is shown in Figure.2. Although not implemented in this study, we could easily remove harmful bot responses by filtering the pre-generated bot utterances. Each $u_t^{'k}$ is transformed into an embedding $e_{u_t^{'k}}$ using sentence-bert [1], and applied in the next matching module.

The matching module measures the cosine similarity $score_k$ between the embedding e_{u_t} of the actual user utterance and each predicted user utterance embedding $e_{u_t^{'k}}$ after user utterance u_t. We describe the maximum similarity among N predictions as $score_{max}$, and the predicted user utterance with the maximum similarity as $u_t^{'max}$. If this maximum similarity exceeds a certain threshold T, the prediction is considered successful and the pre-generated bot utterance $b_{t+1}^{'max}$ is returned as the response. If $score_{max}$ is below T, the generation is performed as usual by the LLM. The only delay during the matching module is the time required to create the embedding of u_t and measure the cosine similarity with the predicted user utterance embeddings. The time for this matching is negligible compared with the time required for the generation by LLM. Thus, if the prediction is successful, it can be finished much faster than response generation by LLM. Even if the prediction fails, the matching time is small compared with the generation time, so the response time is almost the same as in existing systems.

[1] https://huggingface.co/sonoisa/sentence-luke-japanese-base-lite

4. Experiments

4.1. Data

We conducted experiments on two datasets to investigate the effectiveness of our system. These datasets are based on actual dialogues between a user and a bot, containing suitable evaluation data for our system. Both datasets are in Japanese; thus, the experiments were conducted in Japanese.

First, we used dialogues collected in the Dialogue Robot Competition 2022 [14] (robot competition). This competition is task-oriented, with the bot playing the role of a travel agency clerk and the user playing the role of a customer. The user interacts with the bot for 5 min to determine the user's travel plan in Odaiba, Tokyo. In this task, the bot takes the initiative in the conversation, often asking the user questions. Therefore, the user's next utterance is likely to be an answer to that question, making the look-ahead task relatively easy. The seven dialogues contain a total of 71 user utterances. We used four utterances to determine the threshold of matching judgment and set the threshold as 0.75.

The robot competition data were recorded as video, meaning that we could measure the speech time of the bot and user. The average speaking time of the bot utterance was 20.77s, and the time including the succeeding user utterance was 23.90s. We note that due to the nature of the task, there were cases where explanations about tourist attractions took more than 30 seconds to complete. When these long utterances were omitted, the average speaking time of the bot was 11.21s and the time including the user utterance was 14.05s. The distribution of the time for short utterances is shown in Figure 3. We can see that most utterances are distributed in the range 5–10s

We also used dialogues collected in the open track of the Dialogue System Live Competition 5 [15] (live competition). In this task, users were instructed to talk about a randomly selected topic, and the bot and user engaged in a chit-chat dialogue under that topic. Unlike the robot competition, this task gives the user control of the conversation, and the bot utterance is often an answer to the user's question. In fact, in robot competition, 91% of the bot utterances are questions, while in live competition the percentage drops to 41%. Thus, in live competition, the user's next utterance may continue talking about the same topic as the bot's utterance or move to a different topic, making it difficult to predict the next utterance. Therefore, the system performance is expected to be worse in this task than in the robot competition. A total of 810 data points were collected from 50 dialogues. We used ten utterances to determine the threshold of matching judgment and set the threshold to 0.80.

Table 1: *The result of our look-ahead system on robot competition.*

Patterns	Duration		Look ahead accuracy		Matching Precision	Naturalness
	look-ahead duration	matching duration	number of matches	MMS		
N=3	3.99	0.0079	15/67	0.59	4.21/5	0.89
N=5	6.29	0.0080	17/67	0.62	3.89/5	0.70
N=10	9.28	0.0081	15/67	0.63	4.11/5	0.78
N=15	12.11	0.0091	19/67	0.63	4.11/5	0.77
N=20	13.6	0.0092	15/67	0.63	3.71/5	0.74

Table 2: *The result of our look-ahead system on live competition.*

Patterns	Duration		Look ahead accuracy		Matching Precision	Naturalness
	look-ahead duration	matching duration	number of matches	MMS		
N=3	3.58	0.0022	59/800	0.51	3.41	0.71
N=5	6.96	0.0022	42/800	0.52	3.33	0.70
N=10	8.37	0.0023	58/800	0.54	3.29	0.71
N=15	9.73	0.0023	44/800	0.53	3.41	0.69
N=20	11.44	0.0024	52/800	0.53	3.52	0.67

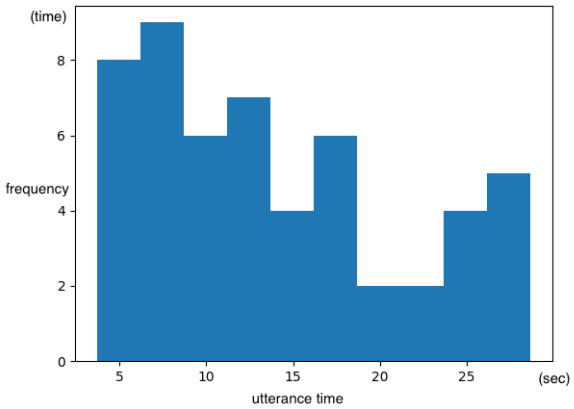

Figure 3: *Distribution of speaking time including user speech*

4.2. Evaluation metrics

We evaluated our system using four metrics. The first indicator is related to the time required for look-ahead and matching. Specifically, we measure the *look-ahead duration*, which is the time taken to predict N user utterances, and the *matching duration*, which is the time taken to match the user's actual utterance with the predicted utterances.

The second indicator relates to the performance of the look-ahead process. We measure how accurately u_t can be predicted based on the history of previous dialogues. This metric is measured by the *number of matches* and the mean maximum similarity (*MMS*). The number of matches refers to the number of successful look-ahead data for which the maximum similarity $score_{max}$ exceeds the threshold, and MMS refers to the average of $score_{max}$ of all data.

The third indicator is the matching precision between predicted and actual user utterances ($u_t^{'k}$ and u_t, respectively). If an inaccurate predicted user utterance is considered to match the actual user utterance, a bot utterance under the wrong context will be returned as a response. To evaluate the accuracy of matching decisions, we annotated the degree to which the predicted utterances whose maximum similarity exceeded a threshold matched the actual utterances. We manually rated the degree of agreement between the predicted and actual utterances on a 5-point scale ranging from 1 (no agreement) to 5 (perfect agreement). We call the average rate the *matching precision*.

Table 3: *Comparison of mean maximum similarity when the bot's last utterance is an question and when it is not.*

	question	not question
robot competition	0.66	0.64
live competition	0.55	0.51

The fourth indicator is the degree to which the pre-generated bot's response aligns with the previous dialogue history. Even when $u_t^{'max}$ accurately predicts u_t, if the next $b_{t+1}^{'max}$ is out of context, the benefit of the look-ahead technique is lost. Therefore, we manually annotate whether $b_{t+1}^{'max}$ is natural, considering the dialogue history, on a 2-point scale (0: unnatural, 1: natural). We call this metric the *naturalness*.

The matching precision and naturalness are subjective evaluations, so we averaged the three annotators' evaluations.

5. Results

Tables 1 and 2 present the evaluation results in robot competition and live competition respectively. First, we examine the response times of our system. The results for the look-ahead duration indicate that each pattern has a generation time of approximately 1s in both settings. Thus, the time required for look-ahead increases as the number of patterns increases. As described in Section 4.1, most utterances are distributed in the range 5–10s; thus, from our result, N=3–5 is considered practical for look-ahead. The time required for matching (matching duration) is almost the same regardless of the number of patterns, and is negligible compared with the time for look-ahead.

The look-ahead accuracy, matching precision, and naturalness are higher for the robot competition than for the live competition, as expected. We now analyze the differences in detail for each evaluation metric.

5.1. Look ahead acuracy

The number of matches in the robot competition data shows that our system achieved high-similarity look-ahead for nearly 20% of utterances (15 out of the 67), indicating the effectiveness of our look-ahead module. However, its success rate dropped to about 5–10% when using the live competition data. The MMS

Table 4: *Two samples of successful look-ahead with high match score and one sample of failed look-ahead*

Dialogue history	Golden User Utterance	Predicted User utterance	max sim
Bot: Have you ever been to either of them before? **User:** I chose to visit because I have never been to both. **Bot:** We want to consider which spot is better for you. Please let us know your information. Who do you go sightseeing with?	I'm going alone.	I'm going alone.	0.999
User: Let' talk about the universe. **Bot:** Universe! Sounds interesting. What would you like to talk about?	If you know, could you please tell me what the beginning of the universe was like?	Tell me about the beginning of the universe.	0.908
Bot: Is it Marco? His video looks so delicious that I tend to make it. **User:** I often watch his videos because I can actually see the steps. **Bot:** If you have any other favorite food, please let me know!	By the way, have you ever seen the Wizard of Oz?	I like sweets	0.269

exhibits a similar tendency, showing that live competition data are, on average, more challenging to predict.

To analyze the performance degradation in live competition, we perform an additional analysis in Table 3 of the MMS when the bot's last utterance is a question and when it is not. The result shows that when the bot's utterance is a question, the look-ahead performance of the following user's utterance is likely to improve. This result is consistent with our expectations in Section 4.1 and indicates one reason for the lower performance scores in live competitions with fewer questions. However, regardless of the question, the overall performance with the live competition data is worse than that with the robot competition data. Compared to the robot competition, the live competition dialogues often fail in turn-taking because the user actively speaks, which often overlaps with the bot's speech. As a result, the look-ahead targets are often user utterances that are shorter than the original sentence (e.g., "But she is") or two utterances combined into one, increasing the prediction difficulty.

The MMS increases slightly as the number of patterns increases, indicating that more diverse utterances can be covered when there are more patterns. However, there is no strong correlation between the number of patterns and the number of matches. Thus, increasing the number of patterns does not increase the number of high-similarity look-ahead predictions.

5.2. Matching precision

The matching precision using the robot competition data was higher than that using the live competition data. With the robot competition data, the score was close to 4 and the matching judgments were basically correct, whereas with the live competition data, the score dropped by about 0.5 and the matching judgments were often incorrect. This drop can be attributed to a combination of two factors. First, the pre-training data used by sentence-bert does not contain many interrogative sentences and are not good at matching between sentences containing questions. As mentioned earlier, user utterances are more likely to be interrogative in the live competition dialogues than in robot competitions because the user takes the initiative. Therefore, our system's matching precision was lower when using the live competition data. We could improve this score by creating a more robust matching judgment model for question sentences.

5.3. Naturalness

We investigate the naturalness of the pre-generated bot's responses. On average, over 80% of the pre-generated bot responses using the robot competition data were judged to be natural. In particular, in the N=3 setting, nearly 90% of the responses were judged to be natural, indicating that even pre-generated responses can take advantage of the context-awareness capabilities of LLMs.

However, using the live competition data, the naturalness dropped to about 70%, because of the lower matching precision performance. Incorrect matching judgment can lead to responses with the incorrect dialogue context. In fact, additional analysis showed that for data with a matching accuracy of 4 or higher, the naturalness using both live and robot competition data exceeded 90%, while for data with a matching accuracy of 2 or lower, the naturalness was less than 30%.

Looking at one dialogue with an unnatural response, the first bot's utterance was "No, I haven't climbed Mt. Fuji" (b_t). Our system predicted that the next user utterance would be "Do you think climbing Mt. Fuji is difficult?" ($u_t^{'}$) and judged this prediction successful against the actual user utterance of "Do you want to climb Mt. Fuji?" (u_t). As a result of judging the wrong predicted utterance as a match, our system returned the unnatural response "Yes, Mt. Fuji can be a bit tough for beginners" ($b_t^{'}$), which has the wrong context. Thus, the matching model should be enhanced to improve the naturalness of our look-ahead system.

5.4. Case analysis

Table 4 presents several evaluation samples from our system. The first sample is a successful look-ahead in a robot competition. The bot's last utterance was the question "Who do you go sightseeing with?", a context in which the user's answer could easily be narrowed down. The user response to this question was a typical answer and led to a high maximum similarity.

The second sample is a successful look-ahead in a live competition. In this sample, the bot asks the user what topics on universe he/she wants to talk about. Although this question is more difficult to predict, our system successfully looks ahead with high similarity, indicating the effectiveness of our method. The final sample is a look-ahead failure in live competition data. The bot is talking about food, but the user is switching to a topic about the Wizard of Oz, making prediction difficult. This sample shows the characteristics of live competitions, where users control the conversation.

6. Conclusion

In this study, we developed a method that reduces the response time in a spoken dialogue system using an LLM by predicting the subsequent user utterance and pre-generating the bot's response. An evaluation of our system showed that the look-ahead module predicts approximately 20% of the user's utterances with a high similarity, which effectively improves the system's response speed. However, incorrect matching judgments of the look-ahead result led to the return of unnatural utterances, so it is essential to improve the matching judgment performance. One future direction of research is to have actual users try dialogue systems equipped with our system. Another direction would be to use the dialogue history to determine whether the system should perform the look-ahead task, potentially reducing the computational cost of our system.

7. References

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Gray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=TG8KACxEON

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[3] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai," 2023.

[4] O. D. Forum, "Gpt-3.5 and gpt-4 api response time measurements," 2023, https://community.openai.com/t/gpt-3-5-and-gpt-4-api-response-time-measurements-fyi/237394.

[5] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, and S. C. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0903616106

[6] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, p. 36–45, jan 1966. [Online]. Available: https://doi.org/10.1145/365153.365168

[7] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 3776–3783.

[8] T. Yamazaki, T. Mizumoto, K. Yoshikawa, M. Ohagi, T. Kawamoto, and T. Sato, "An open-domain avatar chatbot by exploiting a large language model," in *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, 2023, pp. 428–432.

[9] Y. Cheng, W. Liu, W. Li, J. Wang, R. Zhao, B. Liu, X. Liang, and Y. Zheng, "Improving multi-turn emotional support dialogue generation with lookahead strategy planning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3014–3026. [Online]. Available: https://aclanthology.org/2022.emnlp-main.195

[10] Y. Kishinami, R. Akama, S. Sato, R. Tokuhisa, J. Suzuki, and K. Inui, "Target-guided open-domain conversation planning," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 660–668. [Online]. Available: https://aclanthology.org/2022.coling-1.55

[11] Z. Jiang, X.-L. Mao, Z. Huang, J. Ma, and S. Li, "Towards end-to-end learning for efficient dialogue agent by modeling looking-ahead ability," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, S. Nakamura, M. Gasic, I. Zuckerman, G. Skantze, M. Nakano, A. Papangelis, S. Ultes, and K. Yoshino, Eds. Stockholm, Sweden: Association for Computational Linguistics, Sep. 2019, pp. 133–142. [Online]. Available: https://aclanthology.org/W19-5918

[12] E. Ben-David, B. Carmeli, and A. Anaby-Tavor, "Improved goal oriented dialogue via utterance generation and look ahead," 2021.

[13] S. Zhang, J. Zhao, P. Wang, Y. Li, Y. Huang, and J. Feng, ""think before you speak": Improving multi-action dialog policy by planning single-action dialogs," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 4510–4516, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/626

[14] R. Higashinaka, T. Minato, H. Nishizaki, and T. Nagai, "Proceedings of the dialogue robot competition 2022," 2022.

[15] R. Higashinaka, T. Takahashi, S. Horiuchi, M. Inaba, S. Sato, K. Funakoshi, M. Komuro, H. Nishikawa, M. Usami, T. Minato, K. Sakai, and T. Funayama, "The dialogue system live competition 5," *The Japanese Society for Artificial Intelligence, SIG-SLUD*, vol. 96, p. 19, 2022.