Construction and Evaluation of Sales and Educational Dialogue Systems in Parallel Conversations

Shota Mochizuki[1], Sanae Yamashita[1], Rio Suzuki[1], and Ryuichiro Higashinaka[1]

***Abstract*— In an avatar-symbiotic society, avatars are tasked with engaging in basic conversations with users, with human operators stepping in to intervene in these conversations when issues arise. This approach, called parallel conversations, is designed to reduce the burden on humans while collaboratively accomplishing tasks in conversations. While such an approach has been previously tested in scenarios involving guidance and casual conversations, it has not been evaluated for more advanced dialogue tasks. In this study, we construct dialogue systems in the domains of sales and education and assess their potential for engaging in parallel conversations.**

I. Introduction

In the context of an avatar-symbiotic society, research has been conducted on parallel conversations using avatars [1]. With parallel conversations, it is necessary to construct 1) a dialogue system that enables avatars to autonomously engage in conversations, 2) a summarization system that allows operators to efficiently understand the content of the conversation before intervening [2], and 3) a control interface for intervention. While parallel conversations have been implemented for tasks such as guidance and casual conversations [3], their feasibility for more advanced tasks has not yet been evaluated. The objective of this study is to test the feasibility of parallel conversations in more advanced tasks such as sales and education.

II. Approach

Figure 1 presents our approach to verify the feasibility of parallel conversations. The approach involves evaluating the autonomous dialogue system, assessing understanding through summaries, and evaluating dialogue control, in that order. In the evaluation of autonomous dialogue systems, we first construct a dialogue system using a large language model and assess task accomplishment in human-system dialogue. In assessing understanding through summaries, we construct a summarization system and compare the created summaries with the original dialogues to evaluate the extent to which the summaries contain the necessary information for understanding the content of the original dialogues. In the evaluation of dialogue control, dialogues are conducted under conditions where either the operator performs all utterances or half of the utterances are managed by the dialogue system using a control interface. The objective is to determine whether using the control interface can lead to effective parallel conversations.

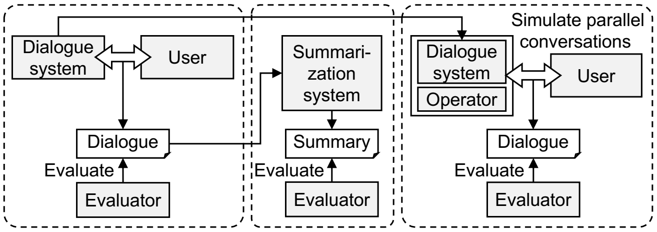

Fig. 1. Approach to verify the feasibility of parallel conversations.

III. Experiment

A. Systems

The dialogue systems were constructed using OpenAI's GPT-4 (gpt-4-1106-preview). For each domain, sales and education, prompts were crafted and the necessary information along with the entire conversation history were input to generate the subsequent system utterance, as detailed below.

1) Sales: The dialogue system operates as a salesperson interacting with users. As a salesperson, it engages with customers who are accessing online shopping sites, listens to the customers' requirements, and suggests the most suitable product from a set of predefined options. In this study, we use smartphones as products.

We prepared two types of dialogue systems: baseline and proposed. For the baseline system, we provided a simple instruction to perform sales along with product information in the prompt. For the proposed system, in addition to the baseline information, we manually provided 30 statements describing useful tips that salespeople could use to improve the quality of their customer service.

2) Education: The dialogue system operates as a teacher interacting with users acting as pupils. The system engages in dialogue to explain and answer questions, thereby helping the user to deepen their understanding of the theme. As themes in this study, we use "Kazuo Ishiguro" and "organ transplantation" in order to cover humanities and scientific content. The system was provided with information related to the respective theme (hand-crafted bullet points summarizing the theme and Q&A pairs, as well as content from Wikipedia articles related to the theme) in the prompt.

B. Summarization System

The summarization system was created using GPT-4 (gpt-4-0613). GPT-4 was given prompts (that same as those used in our prior work [4]) to generate a summary from the input dialogue.

Nine variations of summaries were created based on summary format, input format, and compression rate. There were

*This work was supported by JST Moonshot R&D Grant (JPMJMS2011).

[1] Graduate School of Informatics, Nagoya University, Japan. mochizuki.shota.k8@s.mail.nagoya-u.ac.jp

two types of summary formats: text format summary (TFS) and dialogue format summary (DFS). TFS is an abstract third-person summary that is commonly used, whereas DFS represents the dialogue as a condensed dialogue [5]. Two types of input formats were considered: sequence organization (SO) and the entire conversation history (all). Sequence organization refers to chains of utterances such as questions and answers. In the case of sequence organization, GPT-4 was used to extract SOs from the input, generate summaries for each chain, and then concatenate these to form a summary of the entire conversation. For 'all', summaries were generated directly from the input dialogue. Compression rates were set at 20%, 30%, and 40%.

C. Evaluation Measures

The dialogue and summarization systems were evaluated subjectively across four categories: task accomplishment, understanding of the dialogue situation, cognitive cost, and effectiveness of the control interface. For task accomplishment, in sales, satisfaction was rated on a five-point scale. In education, ten aspects, including knowledge gain, satisfaction, and factfulness, were rated on a seven-point scale Cognitive cost was assessed using the mental workload scale of NASA-TLX [6]. Regarding the effectiveness of the control interface, in sales, smoothness, comfort, clarity, appropriateness of questions, and satisfaction were rated on a five-point scale. In education, the same ten items used for task accomplishment were rated on a seven-point scale.

D. Control Interface

We built a simple web-based control interface on which the performance of parallel conversations can be tested. On this interface, the user can engage in text-based dialogue with the system or the operator; here, the ratio of operator and system utterances can be controlled. We prepared two conditions: one where the system managed half of the utterances and one where the operator managed all utterances.

E. Conducting Experiments

In sales, for both the baseline and proposed systems, we collected 20 dialogues each with users to evaluate task accomplishment. In the education domain, for each theme, we collected 10 dialogues with users to evaluate task accomplishment. Regarding the summarization system, for each domain, nine types of summaries were generated for the 20 dialogues conducted between the dialogue system (the proposed system in the case of sales) and users, resulting in a total of 180 summaries for each domain. The evaluation using the control interface was done with 20 users in the sales domain and 17 users in the education domain.

IV. Results

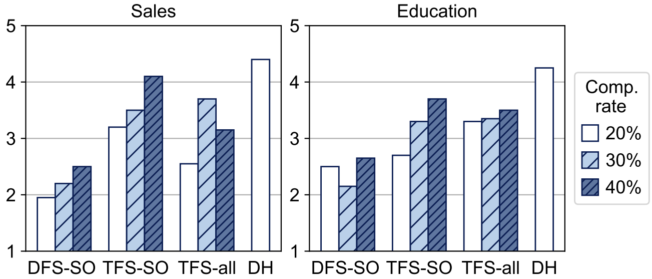

Fig. 2. Results for cognitive cost. The lower the better.

Regarding task accomplishment, in the sales domain, the majority of dialogues using the proposed method were rated with the highest score of 5, achieving a high task accomplishment rate. Only about half of the dialogues were rated as 5 in the baseline, thus suggesting the usefulness of including strategies in the prompt. In education, most items scored above 5 out of 7, indicating a high task accomplishment rate in educational dialogues.

Regarding understanding of the dialogue situation using the summaries, in both domains and for all summary formats, summaries of 20%, 30%, and 40% compression rates enabled understanding of about 50%, 60%, and 70% of the original dialogue content, respectively.

Figure 2 presents the results for cognitive cost using the control interface. DH, representing the entire dialogue history, naturally had the highest cognitive cost. In both domains, DFS-SO managed to reduce the cognitive cost to about 40% to 50% of DH, whereas TFS-SO and TFS-all did not reduce the cognitive cost as much. This suggests that the DFS format reduces cognitive costs.

Regarding the effectiveness of the control interface, in sales, the condition where the system managed half of the utterances outperformed the condition where the operator managed all utterances in four items (excluding smoothness). In education, the half condition outperformed in eight items. This indicates that in both domains, using the control interface to let the dialogue system manage half of the utterances can achieve dialogue quality comparable to that of human-to-human dialogues.

V. Conclusion

In this work, we constructed dialogue and summarization systems to verify the feasibility of parallel conversations in the sales and education domains. The results demonstrate the feasibility of implementing parallel conversations.

References

[1] T. Kawahara, N. Muramatsu, K. Yamamoto, D. Lala, and K. Inoue, "Semi-autonomous avatar enabling unconstrained parallel conversations–seamless hybrid of WOZ and autonomous dialogue systems–," *Advanced Robotics*, pp. 1–7, 2021.

[2] S. Yamashita and R. Higashinaka, "Optimal summaries for enabling a smooth handover in chat-oriented dialogue," in *Proceedings of ACL-IJCNLP Student Research Workshop*, p. 25–31, 2022.

[3] S. Mochizuki, S. Yamashita, K. Kawasaki, R. Yuasa, T. Kubota, K. Ogawa, J. Baba, and R. Higashinaka, "Investigating the intervention in parallel conversations," in *Proceedings of HAI*, p. 30–38, 2023.

[4] S. Yamashita and R. Higashinaka, "Multifaceted evaluation of automatically generated dialogue format summary," in *Proceedings of IWSDS*, 2024.

[5] S. Yamashita and R. Higashinaka, "Clarifying characteristics of dialogue summary in dialogue format," in *Proceedings of IWSDS*, 2023.

[6] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in psychology*, vol. 52, pp. 139–183, Elsevier, 1988.