Open-Domain Dialogue Management Framework across Multiple Device for Long-Term Interaction

Shin Katayama[1][0000−0002−5614−4412], Nozomi Hayashida[1], Kenta Urano[1],
Takuro Yonezawa[1], and Nobuo Kawaguchi[1]

Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya,
Aichi, Japan
`{shinsan, linda}@ucl.nuee.nagoya-u.ac.jp`
`{urano,takuro,kawaguti}@nagoya-u.jp`

Abstract. This study explores the feasibility of dialogue systems with individuality capable of providing continuous and lasting assistance via a multiple device dialogue system. A framework has been devised to manage dialogue history, allowing for the use of a singular identity across various interfaces, including chatbots and virtual avatars. This framework can summarize and save the dialogue history, which can be utilized to generate responses. The impact of dialogue history sharing on users' interactions with a particular character across various devices was assessed for naturalness, continuity, and reliability. The results indicate that dialogue history sharing can foster more natural and continuous conversations, thereby enhancing the potential for long-term support. This research advances the proposition that a digital agent endowed with a consistent identity across multiple devices can provide personalized and sustained support to users.

Keywords: Dialogue Systems · Dialogue Generation · Dialogue Summarization.

1 Introduction

In modern society, the ownership of multiple information terminals by individuals has become a noticeable trend. This can be observed through the common ownership of a variety of devices including smartphones, PCs, wearables, smart speakers, head-mounted displays (HMDs), and robots by users. Natural and human-like dialogue systems have extensive applications in mental health care and education, as they can act as conversational partners instead of humans. However, current dialogue systems on each device operate with distinct identities presenting a challenge for achieving sustained, long-term interaction with a dialogue system, given time and location constraints. Conversely, human-to-human conversations foster enduring relationships through face-to-face communication, online chats, or video calls, facilitating interaction as the same individual and building social connections through continuous engagement. To address

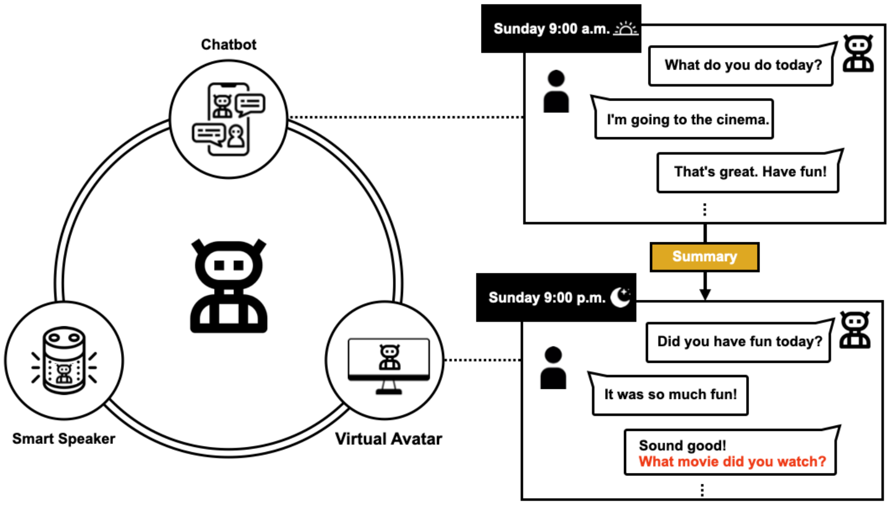


Fig. 1. The concept of dialogue framework across multiple devices.

this challenge, this study proposes a dialogue framework that enables sharing a singular identity across multiple devices, such as smartphones, smart speakers, and robots, by summarizing and sharing dialogue history and generating responses incorporating this information. Figure 1 illustrates our concept. Sharing dialogue history among devices can promote sustained relationships and provide ongoing support. The study employs user experiments to evaluate the effect of sharing dialogue history across multiple devices and investigates the following research questions:

1. Can users perceive a consistent identity when interacting with the system across different devices?
2. Can sharing dialogue history enable more natural dialogues?
3. What information is critical for maintaining social relationships?

2 Related Research

2.1 Agent Migration

Previous research has investigated interactive systems with identities across multiple devices. Early studies proposed agent migration methods [7] that allow an agent to transition between a mobile PC and a self-contained mobile robot. Subsequent investigations have explored agents with personalities across different forms and interactions with virtual characters in human-computer interaction and human-robot interaction [8, 1, 6]. Ogawa et al. [11] proposed the ITACO system, a migratable agent that can move between robots, table lamps, and other entities and build emotional relationships between interactive systems and humans. Gomes et al. [4] developed a migration system to move an artificial pet between a virtual presence in a smartphone and a physical robot, demonstrating that people feel closer to an artificial pet. Recent research [16] has also investigated users' emotional responses toward conversational AI agent migration. However, there remains a need for an established method of agent migration that can move and share dialogue between different devices and modalities using an open domain dialogue history. Previous studies, such as [15, 5], have highlighted the importance of dialogue history content in users' impressions of long-term interactions based on interaction history with companions, environment, and users. Therefore, this study addresses this issue by proposing sharing dialogue history in a conversational system with identities across multiple devices.

2.2 Dialogue Generation for Long-Term Interaction

Using neural network-based language modeling, such as response generation with sequence-to-sequence architecture [14] and Transformer model [17], has become prevalent in the field of natural language processing for dialogue system response generation. Although current open-domain dialogue systems cannot fully replicate smooth and accurate conversations like humans, advances in machine and deep learning techniques have enabled human-like conversations under limited conditions. However, maintaining consistency in personality is challenging for these generative methods. Therefore, research has focused on persona dialogue tasks, which aim to generate consistent responses in line with the persona by incorporating profile information into neural dialogue generation [19, 20]. Existing approaches to persona dialogue systems attempt to incorporate several phrases as an explicit system profile. Consistent response methods have been proposed using speaker identification models [12] and long-term persona memory extraction and continuous updating methods [18]. In addition to having a persona for long-term interaction, considering past conversation history to maintain consistency in dialogue is also essential for enhancing identity. To construct a personalized chatbot, research has proposed methods to automatically learn implicit user profiles from large-scale user dialogue history [13], extract user profiles from dialogue history [21], and summarize dialogue history [2]. Therefore, we propose summarization in this study to share dialogue history between multiple devices. Given the vast amount of dialogue history, it is challenging to consider all of it. Architectures with fixed input lengths limit the length of inputs during response generation, making it impossible to address this issue. In this study, dialogue history is summarized by topic and managed to enable consideration of the summarized dialogue history in new dialogue sessions.

3 Methodology

This study evaluates the hypothesis that is sharing dialogue history among multiple devices will result in more natural dialogue and a sense of identification with the system. In this section, we construct a dialogue management framework for this evaluation and describe the design and implementation of a user study.

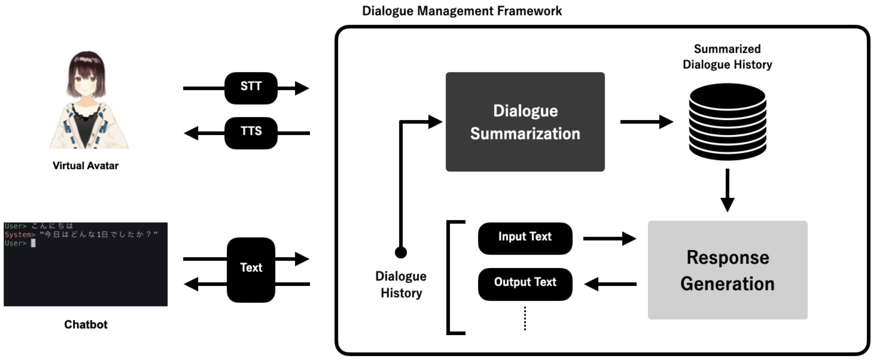
3.1 System Overview

Purpose Traditional dialogue systems face the challenge of limited location and timing for each dialogue interface, which prevents the system from achieving continuous and long-term interaction. In addition, since each interface has a separate identity, dialogue history is not shared. We develop a prototype dialogue management framework to share dialogue history across multiple interfaces to address this problem.

Design As a potential use case for the dialogue system in this study, sharing dialogue history among multiple devices allows continuous dialogue, even over time. Therefore, this study adopts two dialogue interfaces, a virtual avatar, and a text chatbot, and conducts an experiment where they communicate as a single identity. The dialogues targeted in this study are open-domain dialogues, and the dialogue sessions in daily life are set when the user wakes up in the morning and goes to bed at night. Users will be asked to converse three pre-set topics: plan to do today, plan to go today, and plan to eat today, and to interact around these in the morning session and around their thoughts on the topic in the night session. Discussing their daily plan and related feelings on multiple devices makes it possible to maintain continuity and achieve a natural dialogue with a virtual avatar or a text chatbot.

3.2 Implementation

We propose a dialogue framework that enables the sharing dialogue history across multiple devices. The framework is designed to provide the necessary functionalities for the experiment and enable participants to have a seamless conversation experience. It can be applied to various devices, such as virtual agents and chatbots, and can summarize the dialogue history and generate responses based on it. The dialogue management framework consists of two parts: the dialogue summarization part and the response generation part. The dialogue summarization part summarizes the dialogue history for each session and saves it by linking it to a topic. For example, if a dialogue session about what to eat today were composed of ten dialogue turns, the dialogue would be summarized in one sentence and used as input for the response generation part. The response generation part generates appropriate responses to the user's input by considering the context and the user's dialogue history, using a generation-based approach with deep learning. Figure 2 shows the system configuration diagram. These two parts are combined using existing methods, but the framework can be flexible by replacing parts with the latest technology. Overall, the dialogue



```
User> こんにちは
System> "今日はどんな1日でしたか？"
User> 
```

Fig. 2. Dialogue management framework in this study.

management framework proposed in this study can improve the quality of long-term dialogue between users and dialogue systems and improve the consistency and coherence of the dialogue.

Dialogue Summarization Part This part summarizes the content of a dialogue session consisting of multiple dialogue turns. In the prototype system for the experiment, we fine-tuned the BertSum model[10] with the wikiHow dataset[9] to enable Japanese text summarization. We chose the wikiHow dataset because it is more similar to everyday text than the news and headline pairs commonly used in summarization datasets.

BertSum model is an extractive summarization, which involves selecting the most important sentences from a document to create a summary. Our approach involves using this model to extract the most important sentences from a dialogue session and use them as a summary. The summary generated by the model is used as a reference for the response generation part, which generates appropriate responses based on the context of the dialogue and the user's input. Overall, the dialogue summarization part of our proposed framework effectively captures the essence of a dialogue session and provides a valuable reference for generating appropriate responses. Using a summarization model fine-tuned on a relevant dataset enables our framework to perform well on Japanese text, which is particularly important for dialogue systems.

Dialogue Generation Part The response generation part aims to generate a response sentence considering the dialogue history. By utilizing the summarized sentences from the dialogue history of previous dialogue sessions based on the topic of the dialogue, we can easily recall the memory of previous sessions and generate consistent response sentences. In this prototype, we divide each dialogue session into three parts. The dialogue topic is predefined and randomly selected.

1. Greeting: At the beginning of the session, we confirm the greeting and check if the system is ready to engage in dialogue.
2. Question: Next, we ask predefined questions related to the randomly selected topic. If we have a dialogue history related to this topic, we generate follow-up questions to explore the topic in more depth.
3. Chatting: After that, the chats are conducted using a generative-based method with deep learning to generate appropriate responses to the user's flexible utterances continuously. The dialogue continues until it is terminated at any given time, making it a single session.

We use a rule-based method for Greetings and Questions and a generative-based method with deep learning to generate appropriate responses for Chatting. We use Japanese GPT-2, rinna [1] as the backbone for response generation. We fine-tuned the model using an open Japanese dialogue corpus[3] and 10,000 dialogue corpus extracted from Twitter. rinna is a state-of-the-art language model that can generate high-quality Japanese text. We fine-tuned the model with a conversational corpus to adapt it to the nuances of Japanese conversation. The model's ability to generate responses based on the context of the dialogue and the user's input enables our framework to generate more natural and human-like dialogue. For the chatting dialogue session, we explicitly consider the dialogue history by inserting the past dialogue context as the head of the input text with the [SEP] token. By combining the dialogue summarization and response generation parts, our proposed framework enables a more natural and continuous dialogue with users, allowing for a more consistent and coherent conversation. Using state-of-the-art language models fine-tuned on relevant datasets ensures that our framework can perform well on Japanese text and be applicable in various dialogue interfaces.

Interface We adopt two types of dialogue interfaces: a text chatbot that operates on a laptop and a virtual avatar displayed on a screen. Text-based dialogue runs on a terminal, while the virtual avatar uses a template provided by VRoid Hub[2], an avatar creation service. We use Open J Talk[3] for Text-to-Speech, and for Speech-to-Text we use Open AI Whisper[4]. We also implement lip-syncing for the avatar's movements during speech.

4 Evaluation

4.1 User Study

We conducted a user study to evaluate the effectiveness of sharing dialogue history across multiple devices. We recruited 12 participants, eight males and

---

[1] https://huggingface.co/rinna/japanese-gpt2-xsmall
[2] https://hub.vroid.com
[3] https://open-jtalk.sp.nitech.ac.jp
[4] https://github.com/openai/whisper

four females aged between 20 and 50. These participants engaged in dialogue with systems under various real-world scenarios in a simulated environment and were exposed to three experimental conditions as follows:

A No consideration of conversation history
B Consideration of only the user's persona
C Consideration of the user's conversation history

Condition A operates as a conventional dialogue system that functions as a separate dialogue. Condition B presents a system that pretends to have an identity by considering the user's persona, which is limited to the user's name in this study. The user's name is remembered by calling out their name during greetings to create the illusion of memory. Condition C is the proposed method that considers the conversation history. Participants participated in the conversation using text chat and an avatar interface for the user survey. As described in the Methods section, dialogues were conducted twice daily, in the morning and at night, using a role-playing technique in which participants imagined a holiday morning and night (Figure 3). The procedure was as follows:

- In the morning session, participants conversed using one of the interfaces. The topic was randomly selected and the conversation ended when the number of exchanges exceeded seven.
- In the night session, participants used the other interface to converse. The conversation ended when the user's responses exceeded seven exchanges.
- A seven-metric questionnaire was administered to assess factors such as naturalness and fluency using a 7-point Likert scale (1 = totally disagree, 7 = totally agree) at the end of the morning and night sessions.

Participants were instructed to talk about three topics: plan to do today, plan to go today, and plan to eat today. For example, in the morning session, a participant converses with a virtual avatar about plan to go today. In the night session, the participant conversed with a chatbot to converse their feelings of the places they visited that day. Each participant followed this procedure for all three conditions, conducting six interactions. The order of topics, conditions, and interfaces was randomly selected to eliminate biases. Finally, each participant was interviewed to evaluate their experience with the system qualitatively. The evaluation metrics and questions for the participants are summarized in Table 1.

4.2 Results

Figure 4 shows the average and standard deviation of the 7-point Likert scale for each evaluation metric under the three experimental conditions, and Figure 5 shows an example of dialogue by an actual participant. As the figure indicates, it is clear that the proposed method, Condition C, exhibits the highest value compared to the other conditions. Although these results did not demonstrate significant differences, they suggest that the proposed system provides a better

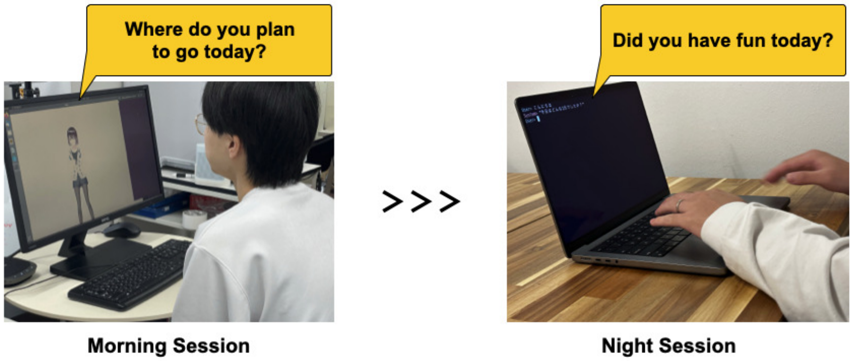


Fig. 3. Morning and night sessions in the user study.

Table 1. Evaluation metrics and questions for a Likert scale.

	Metrics	Question
Q1	Naturalness	Were the system's spoken words phrased naturally?
Q2	Satisfaction	Were the system's conversation satisfactory?
Q3	Comprehension	Did the system accurately understand what you were saying?
Q4	Continuity	Did the system's conversation feel continuous and fluid?
Q5	Relevance	Did you find the conversation with the system trustworthy?
Q6	Continuity	Did the system recall the previous topics of conversation?
Q7	Identity	Did the system have a distinctive personality or character?

dialogue experience regarding satisfaction and continuity, highlighting the system's superiority. In addition, the 12 participants were divided into two groups where six participants used the virtual agent interface in the morning session, while the remaining six used the chatbot interface. Table 2 shows the results for each interface order. The results also demonstrate that Condition C has the highest mean value. This suggests that the evaluation results remain the same depending on the interface usage order. The qualitative interview with the participants provides additional insights into these results. Overall, sharing dialogue history among multiple devices was well received by the participants. However, the participants who assumed that dialogue history sharing was a natural feature were less likely to have a positive experience. For example, during the night session, the effectiveness of the proposed summarization method was diminished when a participant provided a detailed account of the morning session and their impressions, making it difficult to discern differences between the experimental conditions. Moreover, retaining the dialogue history and user preferences may lead to more consistent conversations. For example, Participant 7 noted that machines are capable of remembering their personality, interests, and preferences, eliminating the need to explain them repeatedly. Conversely, Participant

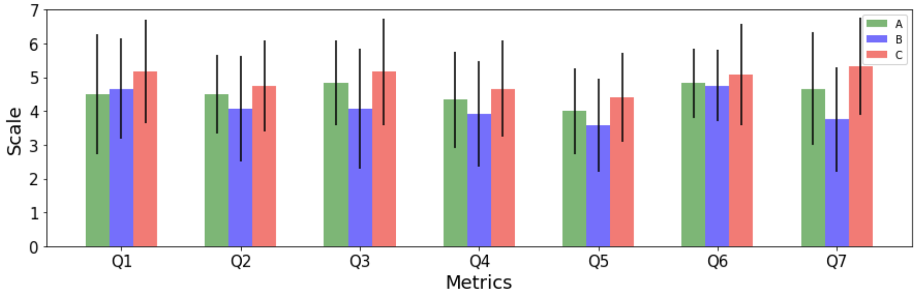


Fig. 4. Results of Likert scale in user study.

5 expressed discomfort at the idea of the system retaining all dialogue history, particularly sensitive information. Meanwhile, Participant 6 proposed that the system should retain significant events and conversations over a week instead of a year. This point of view underscores the need to explore whether the significance of long-term memory varies according to the time frame.

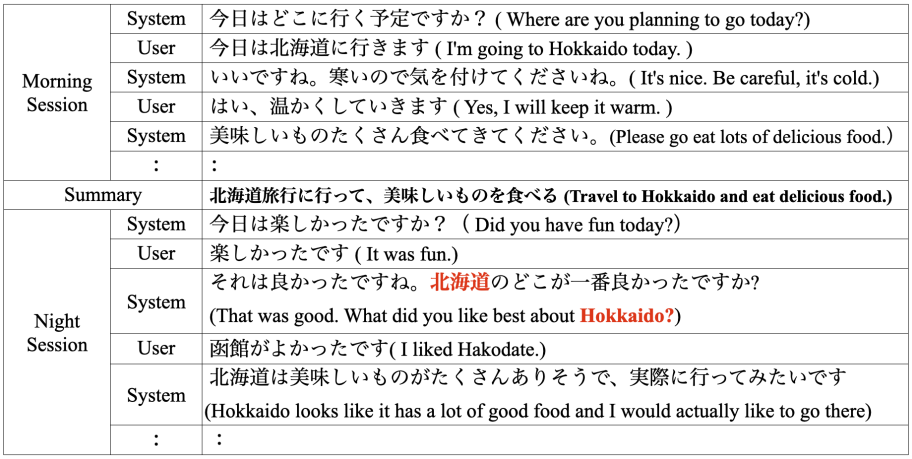
Table 2. Average of Likert scale scores in the user study by order of use of the interface.

	Condition	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Morning: Chatbot Night: Virtual Agent	A	4.66	4.66	4.50	4.16	3.83	4.50	4.50
	B	4.66	4.33	3.83	4.00	3.66	4.50	2.66
	C	**5.33**	**5.16**	**5.16**	**4.83**	**4.33**	**4.83**	**4.66**
Morning: Virtual Agent Night: Chatbot	A	4.33	4.33	5.16	4.50	4.16	5.16	4.83
	B	4.66	3.83	4.33	3.83	3.50	5.00	4.83
	C	**5.00**	**4.33**	**5.16**	**4.50**	**4.50**	**5.33**	**6.00**

4.3 Limitation

Sample size The number of participants in the experiment may have been insufficient to achieve statistical significance. These findings suggest that a larger sample size may be required to increase statistical reliability and generalizability of the results.

System design The experiment design may have been problematic. For instance, if the selected dialogue systems were not sufficiently different, or if the chosen evaluation metrics were not suitable, this could impact the findings. Re-evaluating the chosen evaluation metrics and increasing their number could help detect significant differences more effectively. Furthermore, during the night session, initiating the conversation with a general question such as "Did you have



Morning Session	System	今日はどこに行く予定ですか？ (Where are you planning to go today?)
	User	今日は北海道に行きます (I'm going to Hokkaido today.)
	System	いいですね。寒いので気を付けてくださいね。(It's nice. Be careful, it's cold.)
	User	はい、温かくしていきます (Yes, I will keep it warm.)
	System	美味しいものたくさん食べてきてください。(Please go eat lots of delicious food.)
	:	:
Summary		**北海道旅行に行って、美味しいものを食べる (Travel to Hokkaido and eat delicious food.)**
Night Session	System	今日は楽しかったですか？ （ Did you have fun today?)
	User	楽しかったです (It was fun.)
	System	それは良かったですね。**北海道**のどこが一番良かったですか? (That was good. What did you like best about **Hokkaido?**)
	User	函館がよかったです(I liked Hakodate.)
	System	北海道は美味しいものがたくさんありそうで、実際に行ってみたいです (Hokkaido looks like it has a lot of good food and I would actually like to go there)
	:	:

Fig. 5. Examples of dialogue in user study.

a good day today?" without specific content in the response may lead to a consistent dialogue with the morning session, which could be a factor contributing to non-significant results depending on the content of the conversation.

Environment This study was conducted in a relatively small and restricted setting, and the results should be interpreted in the context of the experimental setting. More longitudinal studies are needed to increase the generalizability of these results. Further experiments conducted in real world environments and long-term memory with dialogue sessions conducted in the actual morning and evening hours would yield better results and more significant findings and consider them as future challenges for this study.

5 Conclusion

In this study, we developed a prototype dialogue management framework that enables sharing dialogue history across multiple interfaces, such as virtual agents and chatbots. The results of our evaluation demonstrated that sharing conversation history across multiple devices led to continuous and natural conversation and could improve the system's consistency. This study provides a new approach to improving the quality of long-term conversations between users and dialogue systems by managing dialogue history. However, due to the small sample size and limitations of the experimental design, statistically significant results were not obtained, and limitations were identified. In future works, we will conduct long-term interaction experiments in real-world environments and utilize more diverse interfaces, such as smart speakers and avatars in virtual reality environments, to identify future challenges and research possibilities in this field.

Acknowledgements.

This research is partially supported by JST CREST(JPMJCR1882, JPMJCR22M4), and NICT(222C0101).

References

1. Arent, K., Kreczmer, B.: Identity of a companion, migrating between robots without common communication modalities: Initial results of vhri study. In: 2013 18th International Conference on Methods & Models in Automation & Robotics (MMAR). pp. 109–114 (2013). https://doi.org/10.1109/MMAR.2013.6669890
2. Bae, S., Kwak, D., Kang, S., Lee, M.Y., Kim, S., Jeong, Y., Kim, H., Lee, S.W., Park, W., Sung, N.: Keep me updated! memory management in long-term conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 3769–3787. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), https://aclanthology.org/2022.findings-emnlp.276
3. Fujimura, I., Chiba, S., Ohso, M.: Lexical and grammatical features of spoken and written japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. In: Proceedings of the VIIth GSCP International Conference. Speech and Corpora. pp. 393–398 (2012)
4. Gomes, P.F., Sardinha, A., Márquez Segura, E., Cramer, H., Paiva, A.: Migration between two embodiments of an artificial pet. International Journal of Humanoid Robotics **11**(01), 1450001 (2014)
5. Grigore, E.C., Pereira, A., Yang, J.J., Zhou, I., Wang, D., Scassellati, B.: Comparing ways to trigger migration between a robot and a virtually embodied character. In: Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8. pp. 839–849. Springer (2016)
6. Ho, W.C., Dautenhahn, K., Lim, M.Y., Vargas, P.A., Aylett, R., Enz, S.: An initial memory model for virtual and robot companions supporting migration and long-term interaction. In: RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication. pp. 277–284 (2009). https://doi.org/10.1109/ROMAN.2009.5326204
7. Imai, M., Ono, T., Etani, T.: Agent migration: communications between a human and robot. In: IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028). vol. 4, pp. 1044–1048 vol.4 (1999). https://doi.org/10.1109/ICSMC.1999.812554
8. Koay, K., Syrdal, D., Dautenhahn, K., Arent, K., Małek, Ł., Kreczmer, B.: Companion migration–initial participants' feedback from a video-based prototyping study. Mixed Reality and Human-Robot Interaction pp. 133–151 (2011)
9. Koupaee, M., Wang, W.Y.: Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305 (2018)
10. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3730–3740 (2019)
11. Ogawa, K., Ono, T.: Itaco: Constructing an emotional relationship between human and robot. In: RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication. pp. 35–40 (2008). https://doi.org/10.1109/ROMAN.2008.4600640

12. Shuster, K., Urbanek, J., Szlam, A., Weston, J.: Am I me or you? state-of-the-art dialogue models cannot maintain an identity. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 2367–2387. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.182, https://aclanthology.org/2022.findings-naacl.182
13. Song, H., Wang, Y., Zhang, K., Zhang, W.N., Liu, T.: BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 167–177. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.14, https://aclanthology.org/2021.acl-long.14
14. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27** (2014)
15. Syrdal, D.S., Koay, K.L., Walters, M.L., Dautenhahn, K.: The boy-robot should bark!-children's impressions of agent migration into diverse embodiments. In: Proceedings: New Frontiers of Human-Robot Interaction, a symposium at AISB. Citeseer (2009)
16. Tejwani, R., Moreno, F., Jeong, S., Won Park, H., Breazeal, C.: Migratable ai: Effect of identity and information migration on users' perception of conversational ai agents. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 877–884 (2020). https://doi.org/10.1109/RO-MAN47096.2020.9223436
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
18. Xu, X., Gou, Z., Wu, W., Niu, Z.Y., Wu, H., Wang, H., Wang, S.: Long time no see! open-domain conversation with long-term persona memory. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 2639–2650. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.207, https://aclanthology.org/2022.findings-acl.207
19. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2204–2213. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1205, https://aclanthology.org/P18-1205
20. Zheng, Y., Zhang, R., Huang, M., Mao, X.: A pre-training based personalized dialogue generation model with persona-sparse data. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9693–9700. No. 05 (2020)
21. Zhong, H., Dou, Z., Zhu, Y., Qian, H., Wen, J.R.: Less is more: Learning to refine dialogue history for personalized dialogue generation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5808–5820. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.426, https://aclanthology.org/2022.naacl-main.426