Koji Inoue

Kyoto University, Kyoto, Japan
inoue.koji.3x@kyoto-u.ac.jp
http://www.sap.ist.i.kyoto-u.ac.jp/
members/inoue/

1 Research interests

The advent of large language models (LLMs) has progressively transformed advanced spoken dialogue systems (SDSs) into a commonplace reality. Expected to be integrated into a wide range of robotics in the future, these systems are poised to be implemented in different societal contexts. The author has heretofore engaged in the realization of several SDSs utilizing android robots (Inoue et al. (2020)). While the functionality of these SDSs has primarily been limited to laboratory settings, future efforts aim to incorporate real-world environments such as hospitals, shopping malls, and schools, thereby exerting a profound societal impact through the advancement of SDS research.

1.1 Social SDSs in real field

While LLMs are powerful tools, they are not guaranteed to handle all social tasks in the real world. Moreover, even with appropriate prompt-tuning, the issue of hallucination can be fatal in social tasks. First of all, it is necessary to organize a taxonomy of various social dialogue tasks through different perspectives. The author's research group has categorized social tasks into two axes: the speaking role and the listening role. For instance, situations that predominantly require the speaking role can be such as "information guide," while situations emphasizing the listening role can be such as "attentive listening." The key point is to design multiple dialogue tasks that evenly cover the space created by these two axes.

To achieve socially capable SDSs, numerous technical aspects must be addressed. For example, the systems must be capable of handling longer dialogues and long-term interactions. Specifically, they need to effectively store and refer to past dialogues as well as the attributes of the interlocutors. While LLMs are based on the transformer architecture, it is important to question whether simply extending the prompt length is sufficient. Human memory mechanisms are more efficient and self-organizing, so it may be worthwhile to explore explicit models inspired by human memory for improved performance. Furthermore, there will be a need for functionality that enables the expression and updating of the system's own personality. By achieving the above-mentioned features, the research goal of the author is to establish a relationship between systems and users through social dialogue, fostering rapport and trust.

1.2 Robust and smooth turn-taking system

When testing SDSs in real-world scenarios, turn-taking always becomes a critical and primitive issue. Conversational robots often face challenges in effectively acquiring turns, leading to situations where the user ends up speaking continuously without interaction. The systems may interrupt and interject in the middle of the user's speech, even before the user has finished their turns. In human-human dialogue, this is not the case owing to a sophisticated mechanism for adaptation, allowing us to engage in conversations with others for extended periods, even several hours. Consequently, conversing with a robot lacking an appropriate turn-taking system can quickly lead to disengagement.

The author has previously proposed several models for turn-taking systems. However, achieving human-level robustness and smoothness in turn-taking still remains a challenge. Additionally, with the emergence of large pre-trained models such as wav2vec 2.0 and AudioLM, there is growing interest in harnessing these models to develop end-to-end systems. Currently staying at KTH Royal Institute of Technology, the author is actively exploring the potential of turn-taking models utilizing large pre-trained models. The ultimate goal is to deploy such models in real-time conversational robots. The mechanism underlying human turn-taking can be seen as a sophisticated architecture that encompasses not only local language understanding but also global dialogue comprehension, response generation, and so on. Ultimately the author aims to investigate models that incorporate these intricate functionalities into SDSs.

1.3 Evaluation method for social SDSs

In the process of practical implementation of SDSs, another crucial aspect is the evaluation methodology. In the field of conversational robots, reliance on subjective evaluations has been common, which poses challenges to research reproducibility and hinders the expansion of the research field. Therefore, efforts are being made to develop objective and effective evaluation metrics. Specifically, the author is working on constructing a framework to indirectly evaluate the "human-likeness" of systems based on users' multimodal behaviors. For example, in human-human dialogues, many interactive backchannels are observed to keep people engaged in the dialogue. Inspired by this, if we observe many backchannels from

Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 24–25
September 11–12, 2023. ©2023 Association for Computational Linguistics

users, it might be said that the system could conduct a conversation in a more human-like manner. The ultimate goal is to empower conversational robots to engage in self-reflection, autonomously learn, and evolve by evaluating their dialogues using objective evaluation metrics.

2 Spoken dialogue system (SDS) research

In the upcoming years, SDS research is expected to shift its focus toward practicality. It is crucial to go beyond mere applications and strive for a more human-like understanding and behavior in SDSs. Note that it would potentially need another discussion on whether human-likeness is needed for SDSs.

2.1 Deeper mind state of user

To achieve a deeper understanding of users, it is essential to conduct studies that delve beyond the surface level of dialogue and explore the inner states of humans. One aspect of the inner state can be identified as *emotion*. Despite the extensive research conducted on emotion recognition and dialogue modeling based on user emotions, there remains a question of whether current models of emotion recognition can adequately capture the intricacies of emotions within dynamically changing social dialogue contexts. In social dialogue scenarios, more subtle emotions undergo dynamic fluctuations. As humans, we adjust our dialogues from micro to macro levels while interpreting these nuanced emotional changes in our conversation partners. By achieving such capabilities, SDSs can explore the user's deeper inner state and become trusted entities in our society.

Furthermore, advancing research in this field will require interdisciplinary approaches that involve fields such as psychology. Therefore, for young researchers and developers in the SDS field, it is desirable to actively acquire not only engineering knowledge but also insights from the humanities and social sciences.

2.2 Relationship with society

Furthermore, for SDSs and conversational robots to truly become social entities, it is necessary for them to engage in not only one-on-one conversations but also in multi-party and multi-session dialogues. However, despite the significance of data-driven approaches in the current era, there is a scarcity of datasets available for learning and simulating such dialogues. Given the unlikelihood of a comprehensive dataset being readily available, it becomes necessary to divide the problem and construct datasets initially for individual issues. For instance, in the context of multi-party dialogues, it is possible to separate the problem into two distinct tasks: multi-party turn-taking prediction and response generation. By repeatedly constructing such datasets and proposing new problem formulations, it becomes essential to solidify the emerging tasks for multi-party SDSs. Furthermore, rather than confining conversations to a single user, it is valuable to aim for situations where information propagates through interactions between the system and multiple users, ultimately fostering a sense of community.

To achieve this, standardization of datasets and experimental systems is necessary. Unlike the presence of a common framework such as ROS (robot operating system) in robotic systems, SDSs often require individual research groups to build their systems from scratch. It would be desirable to develop a common system that includes available datasets to improve this situation.

3 Suggested topics for discussion

The author would like to propose the following topics for discussion.

- What practical and societal dialogue tasks can be achieved with LLM in the coming years?
- To what extent can SDSs delve into the user's inner states? Additionally, how can we ensure the accuracy and reproducibility of SDSs?
- What type of relationship between SDSs and users should be considered ideal for advancing research and development? Should SDSs be convenient tools such as other generative AIs, providing surface-level interactions? Or should they aim for a socially engaged relationship, similar to a friend, where personal matters can be shared?

Acknowledgments

This work was supported by JST ACT-X (JPMJAX2103). The author deeply appreciates the constructive comments and valuable feedback provided by the reviewers.

References

Koji Inoue et al. 2020. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *SIGDIAL*.

Biographical sketch

Koji Inoue received his Ph.D. from the Graduate School of Informatics, Kyoto University, Japan, in 2018. Currently, he serves as an assistant professor at Kyoto University. Additionally, he is currently a visiting researcher at KTH Royal Institute of Technology, Sweden. He has developed a spoken dialogue system for android ERICA. He is a winner of NETEXPLO Innovation 2022 Award.