The Day a System Becomes a Conversation Partner—Exploring New Horizons in Social Dialogue Systems with Large-scale Deep Learning

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, and Hideharu Nakajima

Abstract

People live their lives by casually talking with others on a daily basis. Such "social" dialogue contributes to building trust among people and satisfying their desire to talk with others. There has been a growing interest in social dialogue systems to satisfy the human desire for chatting with others, and we have been working on a wide range of research projects to develop such systems. With the rapid progress in deep learning, high-performance social dialogue systems using deep learning have been proposed. In this article, we introduce NTT's social dialogue system using the latest deep-learning models as well as the current achievements obtained and challenges with this system.

Keywords: social dialogue system, large-scale deep learning, context understanding

1. Question-answering-style social dialogue system

NTT developed a social dialogue system to satisfy people's desire for casually chatting with others. Since interaction with the system is not restricted by location or time, it can increase communication opportunities for people who live alone or in elderly facilities or hospitals. Since we do not need to hesitate to talk to the system, unlike human interlocutors, it is considered promising for a variety of scenarios, such as enabling people who are not good at interpersonal communication to talk with the system or making it easy for a person who is afraid to talk to others about personal issues.

To develop such a dialogue system, we first focused on the characteristics of casual dialogue. The first characteristic is open-domain dialogue, meaning a wide range of topics that can appear in dialogue. We examined the 3680 text of social dialogues we collected from crowd-workers and found that 18,000 topics were included in these dialogues. Even the most common topic, "travel," accounted for only 0.7% of the total, making for an extremely long-tailed distribution. The second characteristic is the diversity of the dialogue flow. In a task-oriented dialogue such as scheduling, there is a certain flow, and a system can be designed with this flow in mind. However, since the flow of a casual dialogue varies greatly from moment to moment, it is difficult to assume the flow in advance.

Because of these characteristics, the initial approach

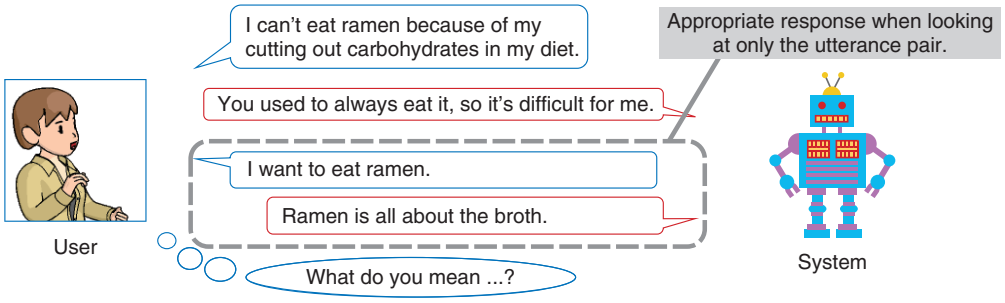


Fig. 1. Problems of a question-answering social dialogue system and NTT's approach.

to developing a social dialogue system was to generate responses to the user's utterances in a question-answer fashion. With this approach, a large number of input-output pairs (response patterns) are prepared in advance, and responses are generated by searching for patterns that are similar to the user utterances. Since a dialogue is composed of a series of utterances, the context of a dialogue should be taken into account. However, the number of possible combinations of utterances is too large to practically generate system utterances while taking into account the context.

There are three typical methods for creating response patterns: rule-based, extraction-based, and generation-based. The rule-based method involves creating response rules. With this method, the rule designer creates a system response to the expected user speech, such as "hello" when the user says "hello," or "good night" when the user says "sleepy." Since the responses are created manually, the system has the advantages of high controllability, low risk of inappropriate speech, and ease of preparing speech that entertains the user such as current events. Because of these advantages, most current commercial social dialogue systems such as Siri are based on rules. One of the disadvantages of rule-based systems is that it is difficult to construct a system that can handle a wide range of topics because the utterances are constructed manually.

The extraction-based method involves extracting and retrieving sentences (examples) from large-scale data and using them in speech. There are two approaches to this. One is to use similar sentences as examples (newspaper articles, blogs, single tweets, etc.), and the other is to use pairs of utterances as examples (dialogue logs, tweet replies, question-answer, etc.). The advantages of this method are that it is inexpensive to implement and can respond to almost any topic. However, the approach that returns similar sentences has the disadvantage that the output tends to be a parroting of the user's speech, while the approach that returns pairs of utterances tends to output utterances with little relevance because the context of the example does not match the context of the current dialogue.

The generation-based method improves the quality of the response utterance while taking advantage of the range of topics with the example-based method. With this method, related topics are extracted from a large amount of text in advance as pairs based on their dependency relations, and a system utterance is generated using the pairs corresponding to the important parts of the user utterance. Therefore, irrelevant sentences and parroting, which are problems with the extraction-based method, can be suppressed, and high-quality utterances can be generated.

NTT combined these methods and took into account their advantages and disadvantages to develop a social dialogue system that generates stable, high-quality responses.

2. Problems with a question-answering social dialogue system and NTT's approach

Despite the improvement of various methods for generating utterance pairs, when we actually talk to a dialogue system constructed with these methods, we sometimes find that the conversation does not mesh well and the dialogue breaks down. The authors investigated such breakdowns and found that there were many responses that were reasonable to individual utterances but were not appropriate for the context of the dialogue. For example, as shown in **Fig. 1**, the system response "Ramen is all about the broth." to the user's utterance "I want to eat ramen." is an appropriate response without looking at the

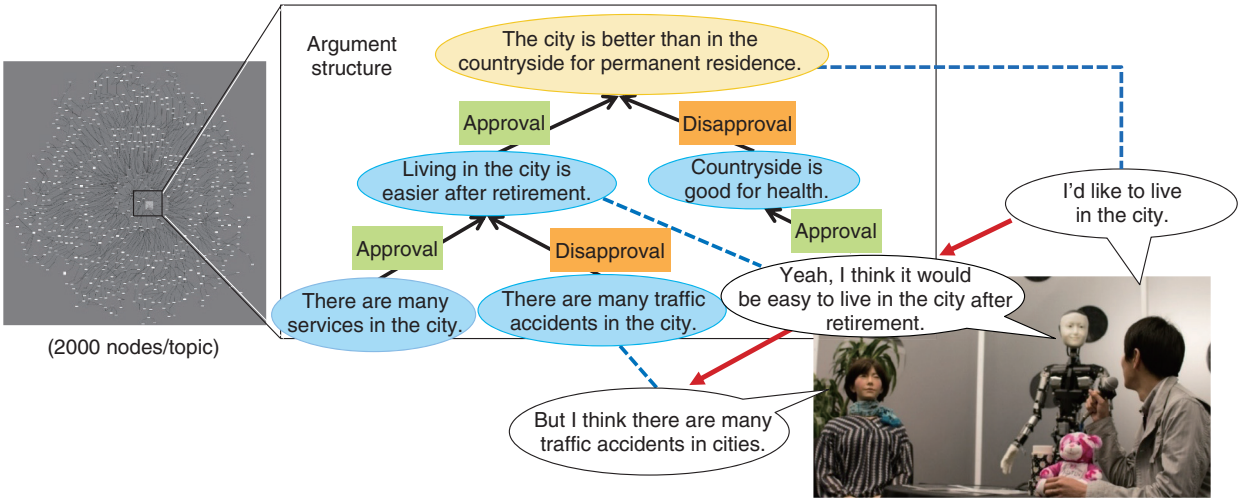


Fig. 2. Discussion dialogue system based on argument structure.

context. However, if the user's utterance is in the context of the user's inability to eat ramen due to carbohydrate restriction, the above system utterance will not make sense because it ignores the user's intention. However, even if we wanted to output the system utterance taking the context into account, it would be impossible with the methods we have discussed thus far due to the huge number of combinations in the utterance history.

NTT took three different approaches to develop a more effective social dialogue system. The first approach is to reset the context by switching speakers. The authors proposed a method of reducing the user's sense of discomfort due to discrepancies with the context, even when using a relatively small amount of data, by having multiple robots collaborate in a dialogue and having one of the robots interrupt and reset the context as needed [1]. We also found that by creating a natural dialogue between robots in advance, we can continue the dialogue naturally by interrupting the inter-robot dialogue when the dialogue is about to break down or create a flow of conversation that develops from the user utterance. By applying these functions, we conducted a demonstration experiment of a "knowledgeable AI (artificial intelligence) robot" at the Kyoto City Zoo and obtained dialogues between visitors and the robot to deepen their knowledge about animals [2].

The second approach is to construct a sequence of dense response patterns restricted to specific topics. Although this is far from the original concept of open-domain dialogue, it is a straightforward approach to consider the context by limiting the topics. However, in a normal social dialogue, we do not know how the topic will develop. For this reason, we focused on discussion dialogues as a middle ground between task-oriented dialogues and social dialogues in which topics can be easily limited. For a particular proposition (e.g., whether to live permanently in the countryside or city), we prepared 20 arguments (e.g., comfort in old age) and constructed a dense sequence of response patterns called an argument structure by connecting opinions supporting and opposing each argument as a tree structure (**Fig. 2**). The developed social dialogue system was presented at the SXSW (South by South West) exhibition in Austin, USA, in collaboration with the Ishiguro Laboratory of Osaka University and ATR (Advanced Telecommunications Research Institute International), to achieve context-based discussion dialogue.

The third approach is to draw the user into a specific context by guiding the user utterances. Through the experiments with the knowledgeable AI robot, we found that even in a social dialogue where there is no obvious flow such as in a task-oriented dialogue, if we can effectively guide the user utterances, we can keep the user in a specific dialogue flow. Using this knowledge, the authors developed a system for chatting about travel using the same design method as the task-oriented dialogue system, with which we can easily define dialogue flow. In an experiment using crowd-sourcing, we confirmed that a very small amount of rules can result in more natural chatting than conventional rule-based and generation-based

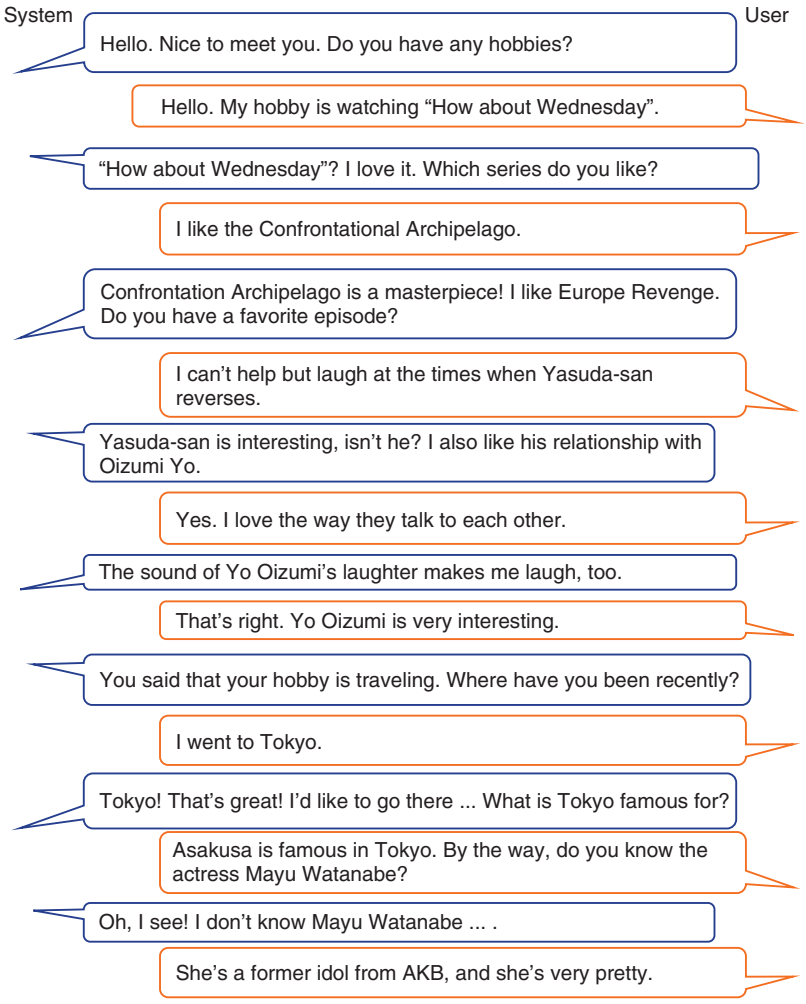


Fig. 3. Dialogue example of Live Competition 3.

systems.

3. Rapid performance improvement with large-scale deep learning

All the systems mentioned thus far either manually select utterances or examples or combine words and apply them to manually constructed templates. There has been rapid progress in deep learning, which is having an enormous impact on natural-language-processing research. In particular, a method called pre-training, with which the naturalness and basic structure of sentences are learned in advance using a large amount of text data, has become important. General-purpose language models trained using this method can achieve very high performance by fine-tuning with a small amount of data for specific purposes such as translation or question-answering.

Social dialogue systems are no exception, and in 2020, a series of high-performance English social dialogue systems based on deep learning were proposed [3]. NTT has also developed a very natural Japanese social dialogue system by pre-training using 2.1 billion utterance pairs collected from Twitter (pairs with the context of several utterances as input and one subsequent utterance as output) and fine-tuning using 200,000 pairs of high-quality dialogue data accumulated in previous research [4]. This system won the top prize in the “Dialogue System Live Competition 3 (Live Competition 3),” a competition of social dialogue systems. To evaluate the system’s ability to handle a wide range of topics, users were required to select two proper nouns as topics and interact with the system to discuss them. **Figure 3**

shows the interaction in Live Competition 3 (system on the left). This user selected a variety show called "How about Wednesday (Suiyou doudesyou)" and the celebrity Mayu Watanabe as topics. It is difficult for conventional systems to respond appropriately to such detailed topics, but the system NTT constructed successfully continued to respond to the user.

4. Future directions

Even though deep learning has made it possible to generate very natural utterances, there are still many challenges. For example, NTT's social dialogue system is trained using only the naturalness of sentences (generative probability) without taking into account the consistency and factuality of the utterances, so it often says things that are inconsistent with past utterances or lies. In addition, it does not remember the content of the dialogue or the other person, so it is difficult to keep repeating the dialogue over a period of several months. We are planning to tackle these issues to develop a higher-performing social dialogue system that continuously satisfies people's desire for dialogue.

References

[1] H. Sugiyama, T. Meguro, Y. Yoshikawa, and J. Yamato, "Improving Dialogue Continuity Using Inter-robot Interaction," Proc. of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 105–112, Nanjing, China, Aug. 2018.

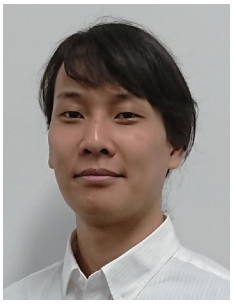
[2] H. Sugiyama, M. Mizukami, and H. Narimatsu, "Continuous Conversation with Two-robot Coordination," Proc. of the 32nd Annual Conference of the Japanese Society for Artificial Intelligence, Kagoshima, Japan, June 2018.

[3] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y. Boureau, and J. Weston, "Recipes for Building an Open-domain Chatbot," Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 300–325, Apr. 2021.

[4] H. Sugiyama, H. Narimatsu, M. Mizukami, T. Arimoto, Y. Chiba, T. Meguro, and H. Nakajima, "Development of Conversational System Talking about Hobby Using Transformer-based Encoder-decoder Model," Proc. of Special Interest Group on Spoken Language Understanding and Dialogue Processing (SIG-SLUD), Vol. B5, No. 02, pp. 104–109, Nov./Dec. 2020.

Hiroaki Sugiyama
Senior Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. in information science and technology from the University of Tokyo in 2007 and 2009 and Ph.D. in engineering from Nara Institute of Science and Technology in 2016. He joined NTT in 2009. He has been engaged in research on chatting dialogue systems for natural human interaction. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), Information Processing Society of Japan (IPSJ), Japanese Society for Artificial Intelligence (JSAI), and Association for Natural Language Processing.



Masahiro Mizukami
Researcher, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E. from Doshisha University, Kyoto, in 2012, and M.S. and Ph.D. in engineering from Nara Institute of Science and Technology, in 2014 and 2017. His research interest includes spoken and natural language processing, especially on non-task-oriented dialogue systems.

Tsunehiro Arimoto
Researcher, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. in engineering from Osaka University in 2013, 2015, and 2018. He joined NTT Communication Science Laboratories in 2018. His research interests include human-robot interaction and dialogue systems.

Hiromi Narimatsu
Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

She received an M.S. and Ph.D. in engineering from the University of Electro-Communications, Tokyo, in 2011 and 2017 and joined NTT in 2011. Her research interests include natural language processing, spoken dialogue systems, and mathematical modeling. She is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), IPSJ, and JSAI.

Yuya Chiba
Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. in engineering from Tohoku University, Miyagi, in 2010, 2012, and 2015. From 2016 to 2020, he was an assistant professor at the Graduate School of Engineering, Tohoku University. He joined NTT Communication Science Laboratories in 2020. His research interests include spoken dialogue systems, multimodal dialogue systems, and human-centric interfaces. He received the IEICE Information and Systems Society Young Researcher's Award in Speech Field in 2014. He is a member of the International Speech and Communication Association, Association for Computational Linguistics, IEICE, and the Acoustical Society of Japan.

Hideharu Nakajima
Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a Ph.D. in science from Waseda University, Tokyo, in 2010. His research interests include prosodic/linguistic/pragmatic analysis of spoken/written messages, spoken language processing (speech recognition, speech synthesis), speech communication with robots/agents, and educational technology.