Chat Dialogue System with Context Understanding

Hiromi Narimatsu, Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, and Noboru Miyazaki

Abstract

Many difficulties arise in developing a dialogue system that can perform conversation in the manner of humans, even for casual conversations. Recent research on chat conversation has led to the development of dialogue systems that can respond to users in a wide range of topics, which was the first major challenge of chat conversation. However, it is still difficult to construct dialogue systems that can properly respond to user utterances according to the dialogue context, and this has often made users feel that the system did not understand what they said. In this article, we introduce our work on a chat dialogue system that has the ability to understand the dialogue context.

Keywords: chat dialogue, context understanding, natural language processing

1. Toward development of a chat dialogue system

Conversations between humans and machines have been increasing with the growing use of agents in smartphones and artificial intelligence (AI) speakers (smart speakers). Most dialogue systems in commercial use are mainly used for executing tasks by giving verbal instructions such as “Call Mr. A” or “Tell me today’s weather,” but there are high expectations for dialogue systems that can chat with humans as a conversational partner. Chatting is said to have many beneficial effects such as helping to organize one’s memory and to improve communication skills. Research has been underway at NTT Communication Science Laboratories on chat dialogue systems from the early stages of dialogue system development.

Unlike with task-oriented dialogue systems, the development of chat dialogue systems is especially challenging because the system must respond to a wide range of topics in user utterances, and the dialogue scenario cannot be designed in advance. With specific tasks such as a restaurant reservation, it is possible to determine in advance the information necessary to make the reservation, such as the date and time or the reserving person’s name and telephone number. In casual conversation, on the other hand, it is impossible to predict the information contained in a user’s utterance. Therefore, it is difficult to make the system respond properly to a variety of user utterances.

Our research group has been working on techniques to develop chat dialogue systems that can respond to utterances in a wide range of topics. One typical technique is to prepare a large number of utterance pairs, such as questions and responses, and use them as training data for machine learning methods. Another technique is to select utterances similar to the user utterance by calculating the similarity between utterances using a dataset of utterance pairs. With the results of previous research, it has become possible to respond to an utterance close to a user’s utterance intention in a one question/one answer format.

However, to make a conversational partner that is more human-like, the system needs to be able to appropriately respond to user utterances according to the context. We introduce here our latest attempts to meet these challenges.

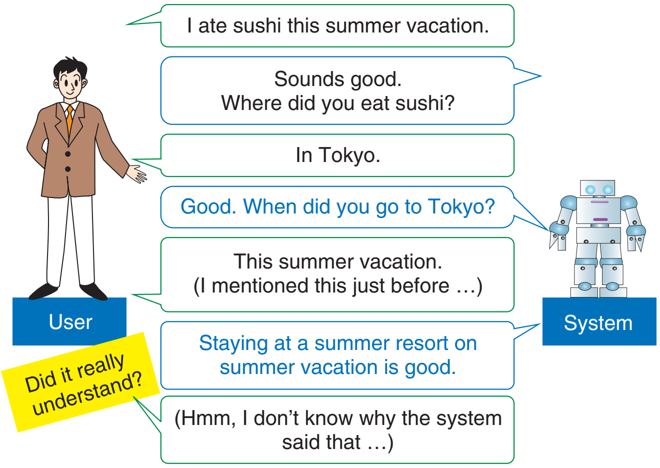

I ate sushi this summer vacation.
Sounds good.
Where did you eat sushi?
In Tokyo.
Good. When did you go to Tokyo?
This summer vacation.
(I mentioned this just before …)
Staying at a summer resort on summer vacation is good.
(Hmm, I don’t know why the system said that …)
User
System
Did it really understand?

Fig. 1. Conversation example between user and conventional dialogue system using utterance pairs.

2. Problems with dialogue systems using single-question/single-answer utterance pairs

In a conventional dialogue system based on utterance pairs, that is, one question and one answer, the approach used is to select a similar response to a user utterance from a large number of prepared utterance pairs [1]. As a result, the system often responds to an aberrational utterance or an utterance that does not match the user’s previous dialogues, and this makes the user feel as if the system does not understand them, even with just a brief dialogue. For example, conversations like that shown in **Fig. 1** are often seen in communication between humans and dialogue systems. In this dialogue, although the user said, “I ate sushi this summer vacation.” in the first utterance, the system asks, “When did you go?” in the fourth utterance. This makes the user think, “I just said ‘this summer vacation,’ but the system didn’t understand ...” Furthermore, the utterance “Staying at a summer resort on summer vacation is good.” suddenly drifts away from the topic of sushi, and it confuses the user, who thinks, “I don’t know why the system said that.”

Such utterances indicating that the response (a) *does not match the dialogue context* and (b) *does not explain why the system said that* may cause users to feel that the system does not understand what they said or that they do not know what the system is trying to say and may thus lead them to give up conversing with the system. Consequently, the dialogue system would be viewed not only as an unskilled conversational system but also a system that does not work as a communication partner with people.

3. Development of a conversational partner

For a dialogue system to at least be recognized as a conversational partner, the problems described in the previous section need to be resolved. The psychologist H. P. Grice also stated a condition for establishing a dialogue, which was to avoid utterances that (a) referred to irrelevant matters (postulate of relevance) or (b) involved unsubstantiated and inappropriate claims (postulate of quality), since these types of utterances lead to the breakdown of dialogue [2]. Therefore, to produce a system that could give substantiated utterances according to the dialogue context while avoiding the above problems, we investigated ways of understanding the dialogue context as well as two utterance-generation methods, that is, utterance generation according to the dialogue context and utterance generation based on evidence. The details are described in the following sections.

4. Understanding the dialogue context

How should a dialogue system understand and maintain context information? We focused on the fact that the user’s experience can often be described using 5W1H (Who, What, When, Where, Why, and How) + impressions, and we considered how understanding could be achieved and how to use the information of 5W1H + impressions as context. The 5W1H framework is very simple, and the simple strategy of asking 5W1H questions is often used in

Table 1. Comparison between location phrases extracted by conventional method and by proposed method.

User utterance (Red: location phrase)	Named entity extractor	Our phrase extractor
I went to Italy this summer vacation.	Italy	Italy
I went to the park near Kyoto Station and saw cherry blossoms.	Kyoto Station	the park near Kyoto Station
I often go to electronics stores.	–	electronics stores

human-human conversations and counseling dialogues. These strategies are often seen in daily life, for example, when we talk about travel or about eating delicious food, the questions “Where did you go?”, “When did you go?”, and “How was it?” are naturally asked in human-human conversations.

How can we develop a system that understands 5W1H + impression information through conversation? Information on time and place, taken from 5W1H information, has been the extraction target in the field of named entity recognition. For example, for the given sentence “I went to Tokyo yesterday,” *yesterday* is extracted as the entity of time, and *Tokyo* as the entity of location. The extraction targets of the named entity recognition are proper nouns and specific expressions of date and time. However, is the information extracted as named entities enough for a system to understand human casual conversation? We examined the phrases that people understand as time or location in actual human conversations and found that phrases other than proper nouns accounted for the majority of location phrases. Specifically, about 70% of location phrases are not named entities.

Therefore, we developed a phrase extractor to extract phrases corresponding to 5W1H + impressions contained in the user’s utterance. We developed the extractor by using the sequence-labeling methods that are effective for named entity recognition. The most representative model is CRF (conditional random field) [3], but methods using deep neural networks have also been proposed recently. First, we manually annotated the words or phrases that people understand as items of 5W1H + impressions to actual conversation between humans. Then we developed the extractor by having it train a model with the annotated conversation dataset [4].

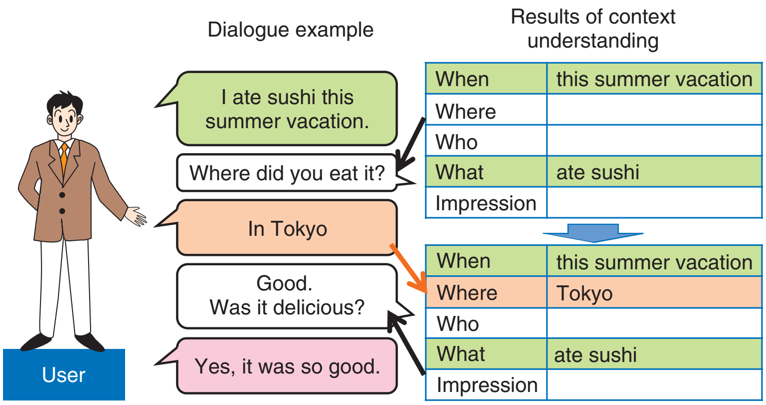
As a result, new types of phrases can be extracted as the target; “the park near Kyoto Station” is extracted as a location, even if it is not a formal proper name, and “I ate sushi” is extracted as a What item. In comparing the results extracted by the conventional named entity extractor and those by our proposed phrase extractor (**Table 1**), we found that phrases including both proper nouns and common nouns could be extracted by our extractor. With this technique, we can develop a system that can understand the context by filling in the 5W1H + impression frames through conversation.

5. Utterance generation aligned with dialogue context

With the results of the contextual understanding described in the previous section, it is easier to generate questions and utterances corresponding to the dialogue context. For example, if the system takes a conversational strategy of asking 5W1H + impressions, this technique prevents the system from asking a question whose answer has already been mentioned by a user (**Fig. 2**). Moreover, this technique helps the system to generate utterances that are appropriately relevant to the dialogue context. For example, if the utterances “I went on a trip during summer vacation” and “I went sightseeing in Tokyo” exist in the dialogue context, our technique helps associate the information of the two utterances and prompts a response such as “Tokyo is hot in summer, isn’t it?” This utterance can be considered more appropriate than the utterance “There is Tokyo Tower in Tokyo,” which is generated by the conventional dialogue system.

6. Utterance generation based on evidence

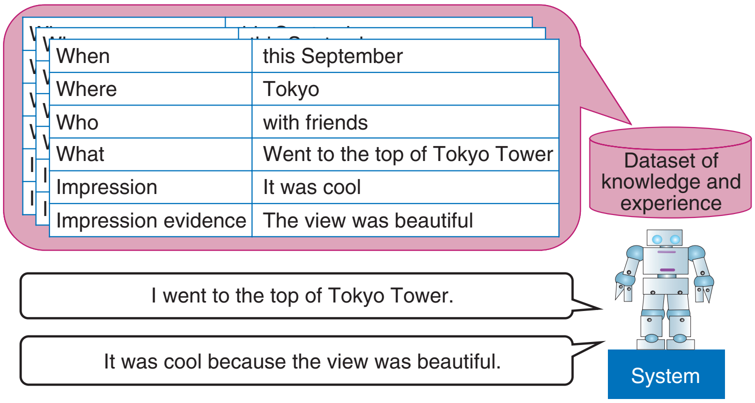
Simply appropriating context does not necessarily result in generation of utterances that show a clear correspondence, or evidence, to why the system produces a particular utterance. Therefore, we proposed a system that provides additional information on the reason the system says the utterance. Here, we introduce an approach using two examples. In the first example, the reason the system asks the question when it does is mentioned. When the system asks, “Can I enjoy it there when I go in summer?”, it provides a reason as supporting information such as “I plan to go there during summer vacation, so I want to



When	this summer vacation
Where	
Who	
What	ate sushi
Impression	

When	this summer vacation
Where	Tokyo
Who	
What	ate sushi
Impression	

Fig. 2. Question generation based on results of context understanding.


Dataset of knowledge and experience
I went to the top of Tokyo Tower.
It was cool because the view was beautiful.
System

When	this September
Where	Tokyo
Who	with friends
What	Went to the top of Tokyo Tower
Impression	It was cool
Impression evidence	The view was beautiful

Fig. 3. Utterance generation based on knowledge and experience of the system.

know this." This additional statement tells the user why the system asked the question [5]. In the second example, the system mentions the reason the system thinks so as the evidence of empathic feelings or impressions when the system expresses such feelings. When the system says, "It was cool," it adds the reason for the impression such as "It was cool because the view was beautiful." We took a simple approach using the structured experience dataset [6], as shown in **Fig. 3** and an utterance template as "I also did [what] and had [impression] because [impression reason]." The utterance "I also went to the top of Tokyo Tower. It was cool because the view was beautiful." is generated by filling each item in the utterance template from an experience dataset. This utterance makes users feel more empathic than the simple utterance, "It was cool." since the system expresses the empathic feelings based on the system's experience and knowledge.

Combining the contextual understanding and context-aligned utterances described previously enables us to add further evidence to context-aligned utterances. This enables the system to produce a dialogue that makes the user think, "This system understands me" (**Fig. 4**).

7. Future work

Through the efforts made in this study, we have developed a dialogue system that is able to understand the context and generate appropriate questions and grounded utterances. This is a major step toward changing an interactive dialogue system that in the past has had users thinking "This system and I do not understand each other" into one that enables them to interact with understanding. If users had a system that

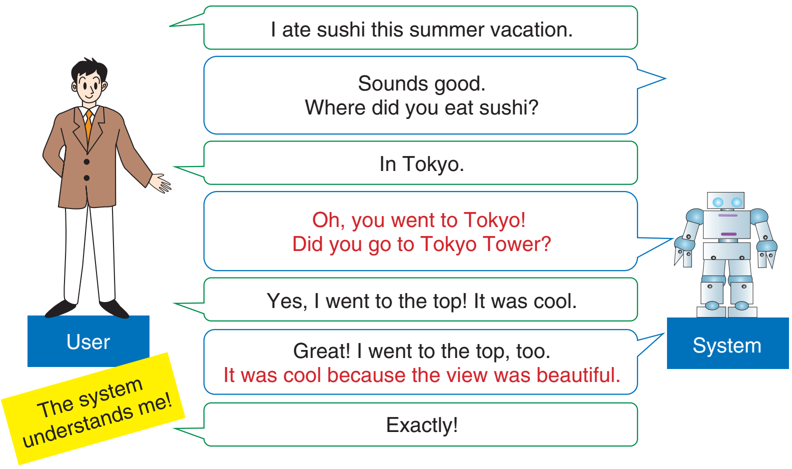

I ate sushi this summer vacation.
Sounds good.
Where did you eat sushi?
In Tokyo.
Oh, you went to Tokyo!
Did you go to Tokyo Tower?
Yes, I went to the top! It was cool.
Great! I went to the top, too.
It was cool because the view was beautiful.
Exactly!
User
System
The system understands me!

Fig. 4. Dialogue example between user and proposed system.

understood what they were saying, they would talk with it in the manner of a conversation between humans. This would also promote the use of dialogue systems in various applications such as communication training and consultation.

However, to achieve this, it is necessary to effectively design the flow of the dialogue and to manually create data that can be used as the knowledge of the system. Moreover, it is not the case that anyone can easily create a similar system. In the future, we will work on a method to automatically generate data through the web or actual conversations between the system and humans, rather than using manually generated data.

References

[1] H. Sugiyama, R. Higashinaka, and T. Meguro, “Towards User-friendly Conversational Systems,” NTT Technical Review, Vol. 14, No. 11, 2016.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201611fa4.html

[2] H. P. Grice, “Logic and Conversation,” Syntax and Semantics, Vol. 3, Speech Acts, P. Cole and J. Morgan (eds.), pp. 41–58, 1975.

[3] J. Lafferty, A. McCallum, and F. C.N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” Proc. of the Eighteenth International Conference on Machine Learning (ICML), pp. 282–289, Williamstown, MA, USA, June 2001.

[4] H. Narimatsu, H. Sugiyama, and M. Mizukami, “Detecting Location-indicating Phrases in User Utterances for Chat-oriented Dialogue Systems,” Proc. of the Fourth Linguistic and Cognitive Approaches to Dialog Agents Workshop (LaCATODA), Stockholm, Sweden, July 2018.

[5] H. Sugiyama, H. Narimatsu, M. Mizukami, and T. Arimoto, “Empirical Study on Domain-specific Conversational Dialogue System Based on Context-aware Utterance Understanding and Generation,” SIG-SLUD, Vol. B5, No. 02, 2018 (in Japanese).

[6] M. Mizukami, H. Sugiyama, and H. Narimatsu, “Event Data Collection for Recent Personal Questions,” Proc. of LaCATODA, Stockholm, Sweden, July 2018.

Hiromi Narimatsu

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

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

Hiroaki Sugiyama

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 a Ph.D. in engineering from Nara Institute of Science and Technology. He joined NTT Communication Science Laboratories in 2009 and studied chat-oriented dialogue systems and language development of human infants. He is a member of IEEE and JSAI.

Masahiro Mizukami

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

He received a Ph.D. in engineering from Nara Institute of Science and Technology in 2017. His research interests include 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 artificial intelligence, human-robot interaction, and dialogue systems.

Noboru Miyazaki

Senior Research Engineer, Cognitive information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.A. and M.E. from Tokyo Institute of Technology in 1995 and 1997. He joined NTT Basic Research Laboratories in 1997. His research interests include speech processing and spoken dialogue systems. He is a member of IEICE, the Acoustical Society of Japan, and JSAI.