

Rationale for Using Chat-Oriented Dialogue System’s Experience to Convey Empathy

Hiromi Narimatsu* Hiroaki Sugiyama Masahiro Mizukami Tsunehiro Arimoto

NTT Communication Science Laboratories

{narimatsu, h.sugi}@ieee.org, {masahiro.mizukami.df, tsunehiro.arimoto.ub}@hco.ntt.co.jp

Abstract

To help create satisfying conversations, a conversational partner must generally empathize with the speaker’s feelings. Although some current dialogue systems can empathize by replying with simple compassionate utterances, such responses lack credibility simply because the system is not human. This paper develops a chat-oriented dialogue system that can convey empathy to users in the manner of a human conversationalist. Our approach focuses on building the system’s efficiency in sharing its own experiences as the origin of empathy. With this basis, the system empathizes with the user’s feelings by sharing the experiences that most resemble those of the user. We developed five types of chat-oriented dialogue systems by changing the method of conveying experience depending on the form of the experience. Our text-chat experiments demonstrate that sharing an experience that resembles the user’s own experiences increases the user’s sense of having received empathy.

1 Introduction

With the spread of agents on smartphones and AI speakers, the opportunities for conversation between humans and machines are increasing. Though most dialogue systems in commercial use are designed mainly for executing tasks such as restaurant recommendations or weather updates, developing non-task dialogue systems such as a conversational agent for the purpose of enjoying conversation has been pursued over the past decade [Zhang *et al.*, 2018]. The elderly, in particular, can be expected to talk with such dialogue agents as partners [Oh *et al.*, 2020].

To make users feel satisfied through conversation, the users’ sense of receiving empathy is essential [Tsang, 2013]. Studies of both human-human and human-machine conversation point out that the user should have the impression of being empathized with when talking with a companion (i.e., “my feelings were understood”) [Cheepen, 1988;

System:	How did you spend this summer vacation?
User:	I went to Paris.
System:	Cool! Did you go to the Eiffel Tower?
User:	Yes. I went there.
System:	Oh, you did. How was that?
User:	It was good. The view is so cool!
System:	<i>It seems it would be good because I found a visitor’s comment saying “the view was beautiful” on the web.</i>
User:	Exactly! ← User feels “the system understands me.”

Figure 1: Conversation example between a user and the proposed dialogue system.

Morris *et al.*, 2018]. Empathy is commonly defined as the ability to identify and understand another person’s situation and feelings [Preece *et al.*, 2001].

For the purpose of developing the ability to empathize in a conversational agent, some studies have been conducted in the field of mental health [DeVault *et al.*, 2014; Morbini *et al.*, 2012; Morris *et al.*, 2018]. One work in the conventional research focused on utterances used to show empathy, and developed a dialogue system that could demonstrate empathy by training the actual text chat messages exchanged between humans [Morris *et al.*, 2018]. Although the system empathizes by trying to imitate human responses, because dialogue often breaks down during conversation between a question-and-answer-based system and humans, users may question the system’s ability to understand, and thus may doubt its ability to empathize. Since such feelings easily inhibit comfortable conversation between users and agents, this problem must be solved. Grice’s Cooperative Principle [Grice, 1975] is illustrative: Not to tell lies in conversation, and to say what you have reason to believe. In order for a system to be accepted as a conversation partner, it must be capable of understanding and produce reliable utterances. Therefore, we need to improve the user’s impression of such abilities of the system.

To tackle this issue, we aim to develop an empathic conversational agent whose expressed empathy appears genuine to the users. In human-human conversation, it has been reported that having a similar experience promotes empathy with another person [Batson *et al.*, 1996; Eklund *et al.*, 2009]. Al-

*Hiromi Narimatsu

NTT Communication Science Laboratories, Kyoto, Japan
Hikaridai 2-4, Seika-cho, Soraku-gun, Kyoto, Japan, 619-0237

though some existing dialogue systems converse through self-disclosure, no previous study has focused on sharing experiences as evidence of holding empathic feelings in the course of developing chat-oriented dialogue systems. Therefore, we focus on sharing similar experiences so that the user perceives the empathy as more believable in the conversational agent, thus making chatting more comfortable.

To show empathy based on the system’s experiences that resemble the user’s experiences, it is essential to establish the system’s own experience, and for the system to be able to draw out and understand the user’s experiences through conversation. In addition, we must consider the method of sharing experiences. Therefore, in this paper, we propose methods to establish the system’s experiences, to understand the users’ experiences through conversation, and to show similar experiences to convey empathy.

Figure 1 shows an example dialogue of our goal. The system asks about the user’s experience and draws out the user’s feelings. The system shows empathy by sharing a system experience similar to the user’s as the system utterance written in italic. The user’s experience is shared not only by a single utterance, but also across multiple utterances, i.e., “I went to the Eiffel Tower” and “The view is so cool.” The system also understands these utterances to correctly select a similar experience. We expect that, with the system expressing such shared experiences, the user will feel that the system actually empathizes with them, even though the user knows the system is not an actual human.

There are two straightforward approaches to describe experience. One is to show evidence based on the system’s own experience such as “I also visited there before, and the city scape was so beautiful.” The other is to show the evidence based on the system’s knowledge, such as hearsay experience. An example of hearsay experience is “It seems good because I found a visitor’s comment saying ‘the view was beautiful’ on the web,” as shown in Figure 1. Both of those descriptions are the same from the viewpoint of showing empathy based on the system’s experience. However, it is possible that the user’s impression is different depending on the way the experience is described. Expressing the experience as the system’s own may appear disingenuous (“there’s no way the system can experience something like this.”) Therefore, we analyze the effects of describing the system’s experience to show its empathic feelings according to the way it is described.

The contribution of this paper is follows:

- Propose a system to convey empathy based on the system’s experience/knowledge that resembles the user’s experience.
- Propose a way to understand the user’s experience through free conversation.
- Show the effects of conveying empathy using the system’s hearsay experience and showing the reason for the impression.

In this paper, we describe our proposed conversational agent that shares the system’s experiences to convey empathy in Section 3 and the ways of expressing empathy using the experiences in Section 4. We develop five types of chat systems by changing the way of showing empathy. We then

evaluate each system from the viewpoint of how users perceive the system’s empathy in Section 5. The evaluation results show that expressing empathy using the system’s experiences as hearsay experiences improves the users’ impression of system empathy.

2 Related Work

In this section, we describe dialogue systems that use strategies to show empathy.

In counseling and mental health, empathy is crucial. Methods based on analysis of human-human conversation have been proposed to show empathy in dialogue systems. In dialogues with robots and virtual agents, human-like voices, facial expressions, and gestures should be made to match the user’s emotions [Fung *et al.*, 2016]. For listening agents, improvement in empathic expressions have been proposed using and controlling the timing of such simple utterances as “good” and “I see.” [Kawahara, 2018]. A method has also been proposed to generate such simple utterances as “sorry” or “good” based on the positive or negative sentiment of utterance contents [DeVault *et al.*, 2014]. Although we can expect these approaches to improve the impression of empathy, they are difficult to apply to non-task oriented conversation because there are specific strategies for counseling and mental health. In addition, these approaches mainly focus on listening, back-channeling, and simple utterances or fixed phrases that can be used, and do not focus on showing evidence of empathy in utterances.

On the other hand, using empathy in chat-oriented dialogue systems is outside mainstream methodology, which has emphasized increasing the naturalness of responses [Yan *et al.*, 2016; Wu *et al.*, 2016] to cover a wide range of topics. Some studies address empathy as a part of dialogue systems. A combination strategy, which repeats a partial phrase of the user’s utterances and simple empathic utterances, was proposed to express empathy and can adapt to a wide range of responses [Higashinaka *et al.*, 2008]. This strategy indicates to the user that “the system understands me.” However, although it is effective to convey empathy, it is easily doubted when dialogue breakdowns occur in conversation. Especially in question-answering based systems for the purpose of covering a wide range of topics, many such dialogue breakdowns occur [Higashinaka *et al.*, 2015].

Another study found that conveying empathy with similar emotions [Tahara *et al.*, 2019] is also effective to convey empathy. However, misfocused emotion or misaligned utterances that the system is possible to occur greatly lower positive impressions toward the system. In the same manner as above, it may increase user skepticism because dialogue breakdowns occur in conversation, or the system cannot answer follow-up questions about the empathy, such as “Where did the system get this emotion?”

Psychological studies in human-human communication suggest that sharing similar experiences improves the impression of empathy [Batson *et al.*, 1996; Eklund *et al.*, 2009]. Although a gender difference was identified in these analysis results, humans empathize with others based on their experiences.

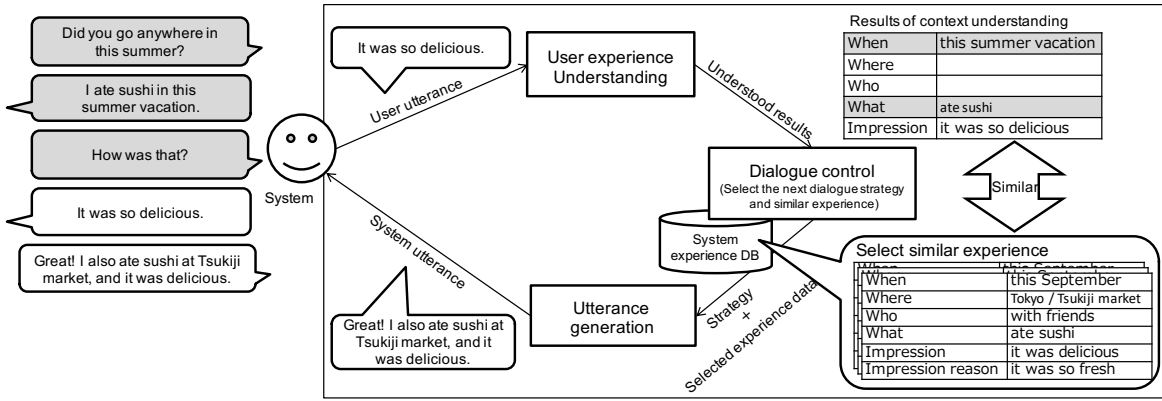


Figure 2: The overall system's operation flow.

In this paper, we aim to increase the user's impression of the system's empathy by providing system experiences as evidence of empathy.

3 Proposed Conversational Agent

To develop a system that can convey empathy based on a system's experience that resembles the user's experience, we developed a pseudo-experience dataset for the system. We then defined that the user's experience extracted through conversation can be expressed by a frame that is the same as the system's experience dataset. First, we describe the overall system flow in Figure 2. The system comprehends the user's experience through conversation and understands the results of the item frames as the user's experience, as shown in the top right table. The system then searches its pseudo-experience database and selects an experience that resembles the user's experience. At that time, it decides the dialogue strategy, such as ask question or convey empathy, depending on the dialogue context. The utterance templates are related to the dialogue strategy, and an utterance is generated using the template and the selected experience.

3.1 Motivation

For using the system's experience as evidence of empathy, we consider two approaches. One expresses it as the system's own experience, such as "I went to the Eiffel Tower and the view is so cool." The other expresses hearsay experience, such as "It seems good because I found a visitor's comment saying 'it was good' on the web." Though both of these descriptions are the same from the viewpoint of showing empathy based on the system's experience, they are expressed differently. To flexibly change the expression of both approaches using an identical system's experience dataset, we adopt template-based utterance generation methods and propose a versatile data design for the system's experience using 5W1H (who, what, where, when, why, how) frames. It is also necessary to elicit and understand the user's experiences through conversation to enable the system to empathize by using its experience that resembles the user's experience. To elicit and understand user experiences, we chose a strategy

that conversationally elicits user experiences based on 5W1H frames [Han *et al.*, 2013]. By making the structure of the system's experience dataset and the understood user experiences the same, it is easier for the system to choose similar experiences.

The base strategy of conversation follows scenario-based dialogue systems that are generally used in task-oriented dialogue systems [Zhang *et al.*, 2020]. However, to apply these strategies to a non-task-oriented dialogue system, it is important to understand what the user says correctly. This study contributes to understanding user experiences through conversation in chatting situations.

The following are its three key components: identifying the predefined system's experience (Section 3.2), understanding and categorizing the 5W1H frames as a user's experiences from her utterances (Section 3.4), and generating empathic utterances based on the system's experience by comparing it with the user's experience (Section 3.5).

3.2 System Experience Dataset

We simply generate system utterances from system experience data by focusing on a template-based method that fills elements in a pre-defined utterance template. We then add Impression and Impression reason to the 5W1H frames. Impression is important to convey empathy and the other 5W1H items, and Impression reason is essential to support the impression as evidence of empathy utterances. The details of the collected data are described in Section 4.2.

To easily generate various types of system utterances, i.e., empathic utterances with the system's own experiences and hearsay experiences, or questions based on the system's experience, we defined the format of the utterance templates and the system experience dataset.

Considering simple system-utterance generation based on utterance templates, we determined in advance what should be written in these dataset frames. For example, when the "I ate sushi in Tokyo" utterance is expected to be generated, we simply make the utterance template "I [what] [where]" and place *ate sushi* in the *what* frame and *in Tokyo* in the *where* frame. Note that the contents of the 5W1H+impression frames are different from the general 5W1H framework. For

System:	Did you go anywhere during summer vacation? (<i>question about user's memory</i>)
User:	Yes. I went to Paris.
System:	Cool! Where did you go in Paris? (<i>simple empathic utterance, follow-up question</i>)
User:	I went to the Eiffel Tower.
System:	Oh, the Eiffel Tower! How was it? (<i>repetition, question about impression</i>)
User:	It was so cool!
System:	<u>Great! I also visited it and the city skyline from the tower was so cool.</u> (<i>simple empathy utterance, system experience</i>)
User:	Yeah, I agree!

Figure 3: Conversational strategy of proposed chat system.

example, the item of *what* includes not only verb but also what to do, such as “ate sushi” and “climb the top of the tower”, and that of *who* includes *with whom* information such as “with friends” and “alone.” Furthermore, we add the category to *what* information such as what to see, activity and food, because we assume to prepare some utterance templates and change them depending on the category.

Moreover, we introduced not only an impression, but also an impression reason to the 5W1H frames to provide a depth of evidence when the system emulates empathy. To mentioning the impression’s reason, we prepared the utterance frame “[Impression] because [impression reason]” and items {impression:it was delicious, impression reason:it was so fresh}. By placing the items in the utterance frame, “It was delicious because it was so fresh” can be generated.

3.3 Conversation Strategy

The conversational strategy used in this study is shown in Figure 3. The following is the overall conversation flow: question about user’s memory → simple empathic utterance, follow-up question based on system’s experience → repetition, and question about impression → empathic utterance based on its experience as reason.

To extend the dialogues for a particular duration, we defined the flow of conversation as a small segment (first asking the items, then asking the impression, and finally conveying empathy), and repeated it recursively by changing the contents (subtopic in conversation) related to the previous context, such as from famous spot to famous food under a traveling topic. Whether or not to ask a question at that time is determined by whether or not the target item of the frame is filled in.

Since the details of utterance generation ways are explained in Section 3.5, the simple strategy of each utterance generation as described below.

Simple Empathy Utterance

We prepared a list of simple utterances, such as “good” or “I see,” as in the conventional studies [Kawahara, 2018; Higashinaka *et al.*, 2008], and an utterance template list for questions and empathy using experiences. A simple empathy utterance is selected based on whether the context is recognized as positive or negative. This recognizer is simply developed by a handcrafted rule, such as “I enjoyed xx” → *positive* or “I couldn’t go xx” → *negative*.

Utterance to show the reason for empathy

To show the reason for empathy, we prepared an utterance

template, such as “[Impression] because [Impression reason]” or “I [What], and [Impression].” “The view from there was so beautiful because I was so tired,” and “I climbed the top of the mountain and the view from there was so beautiful” are the example utterances by filling the templates using the experience dataset.

Follow-up question

We also prepared a list of follow-up questions, which not only include such 5W1H questions as “Where did you go?” and “When did you go?,” but also such factual inquiries as “Did you visit the Eiffel Tower?” and “Did you eat sushi?” In fact, we used the question template: “Did you visit [where]?” or “Did you [what]?”

3.4 Understanding User Experience

The system detects the 5W1H information in the user utterances to identify and understand the user’s experience, and dialogue is conducted to fill the 5W1H frames by choosing questions that were undetected in the dialogue context shown in Figure 4.

Although we use a strategy of asking questions based on the framework of the 5W1H + impression information, we must understand which word or phrase corresponds to which 5W1H + impression. For example, when the system’s question is “Where did you go?,” a simple answer is “I went to the Eiffel Tower.” However, some may not reply simply to the question, but instead mention food, such as “I ate delicious steak.” Other users may instead reply to the question with more details, such as “I went to the Eiffel Tower and enjoyed the view from the top.” In this case, the system needs to understand *what* and *impression*.

Therefore, we developed a 5W1H + impression recognizer. From the 5W1H information, time and place information are the recognition targets in the field of named entity (NE) recognition [Nadeau and Sekine, 2007]. For example, when the given sentence is “I went to Tokyo yesterday,” “yesterday” is extracted as the entity of time and “Tokyo” as the entity of location. The recognition targets of the NE recognition are proper nouns and specific expressions of date and time.

However, the information extracted as named entities is insufficient for understanding casual conversations. In conventional research, the phrases that people understand as time or location in actual conversations between humans are identified, and such phrases (other than proper nouns) accounted for the majority of location phrases [Narimatsu *et al.*, 2018].

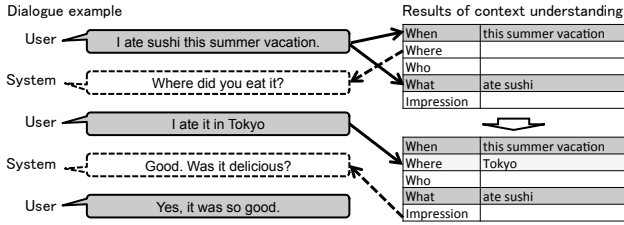


Figure 4: Dialogue example based on results of context understanding.

Therefore, we developed a phrase recognizer to extract phrases corresponding to 5W1H with impressions contained in user utterances. We developed this using sequence-labeling methods that are effective for NE recognition. The most representative model is Conditional Random Fields (CRF) [McCallum and Li, 2003], but, recently, methods using Bidirectional Encoder Representations from Transformers (BERT) [Devlin *et al.*, 2018] have also been proposed, and the performance is high by fine-tuning with small dataset instead of large dataset. The dataset for training the phrase recognizer is automatically generated by simple rules using a system-experience dataset. If the template sentence “I [*what*] [*where*] [*when*]” is given, the sentence “I ate sushi in Tokyo this September” can be generated by filling in each item from the example system-experience dataset in Figure 5. If the *what* item is the extraction target, Begin, Inside, Outside (BIO) tags are given to the generated sentence: {O B-WHAT I-WHAT O O O O}. –WHAT is used to explain the BIO tags for WHAT. We made eight possible template patterns, for example, “I [*what*] [*who*],” “I [*what*]. [*impression*],” and so on, to establish a training dataset.

3.5 System-Utterance Generation Using Similar Experiences

To express the system’s experience that resembles the user’s experience, we calculated the similarity between the user’s experience extracted from the conversation and each system experience. The experience is selected if all the contents of the 5W1H + impression are identical in the user’s and system’s experiences. Unfortunately, no dataset can exactly match the user’s experience. Therefore, we prioritized the matched contents between the user’s and system’s experiences. We gave higher priority to the contents of *what* and *where* for calculating the similarity of experience. We also provided a constraint that at least either *what* or *where* must be extracted from the user’s experience and must be similar to the system’s experience. A simple way of judging similarity is exact matching. For practical use, we introduced a method of partial matching. For example, since there are not many amusement parks in a single area, one of them is extracted from the target area based on the context. If the user does not mention a particular park by name, “I visited an amusement park,” we select a popular amusement park in the area.

By using the selected system’s experience that resembles the user’s experience, the system automatically places the experience data into the utterance template, as described in Fig-

ure 5. The examples of ways to filling the prepared utterance templates with the system experience database are shown in the figure.

4 Experiment

To evaluate the proposed system according to the methods of expressing empathy, we developed five types of chat systems. This section describes the comparison methods (Section 4.1), the preparation of the system’s experience/knowledge dataset (Section 4.2), and the experiment procedure (Section 4.3).

4.1 Baselines and Method Settings

We considered ways of showing experience to express empathy from the viewpoint of experience types. One is simple empathy, as in conventional dialogue systems [Kawahara, 2018; Higashinaka *et al.*, 2008]. The others use a proposed corpus. We assumed two types of experiences: one’s own experience and hearsay experience. In addition, we assumed that directly explaining why a feeling was evoked is another effective way to express empathy from the viewpoint of the depth level of evidence. First, we compared the effects of sharing experiences by changing the types of experiences (one’s own and hearsay) shown in Table 1, and then compared the effects of explaining why the feeling surfaced using an impression reason shown in Table 2. The baseline system in Table 1 is the same as that in Table 2.

Methods	General facts	Experience	Knowledge
Baseline	✓		
OwnExp		✓	
HearsayExp			✓

Table 1: Compared system conditions for evaluation of sharing experiences

Methods	General facts	Experience	+Impression Reason
Baseline	✓		
Experience		✓	
Exp+Reason		✓	✓

Table 2: Compared system conditions for evaluation of mentioning impression reason when sharing experiences

The **baseline** method simply empathizes with the user through utterance repetition, as in conventional dialogue systems [Kawahara, 2018; Higashinaka *et al.*, 2008]. For example, if the utterance “I saw Sagrada Familia” is given, the system partially repeats the user’s utterance as “Oh, Sagrada Familia!” and adds a simple empathy utterance: “That’s good!” or “I see.” In addition, such facts or opinions as “Sagrada Familia is famous in Spain” are added to restrain the impact of the difference in the number of system’s utterances. The completed utterance becomes “Oh, Sagrada Familia! That’s good! Sagrada Familia is really famous in Spain.” We prepared five types of simple empathy utterances, one of which was chosen based on the positive or negative sentiment at the time. Since the effects of simple empathy utterances have already been shown in previous research

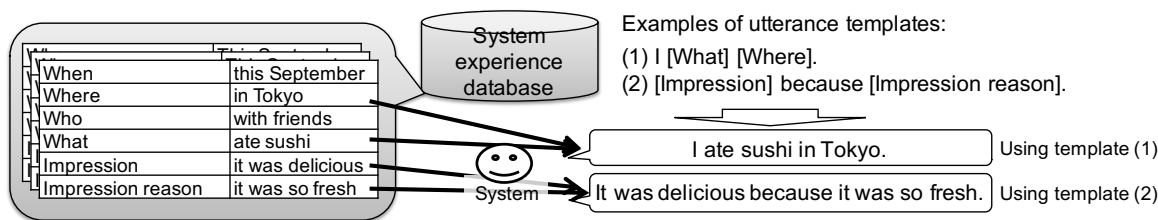


Figure 5: Utterance generation based on results of context understanding.

[Higashinaka *et al.*, 2008], we used the strategy of including simple empathy utterances as the baseline for this evaluation.

OwnExp chooses a system experience that is similar to the user’s experience and mentions this experience and impression. For example, if the user says, “I saw Sagrada Familia,” the system chooses experience data and generates the following: “Oh, I went to Sagrada Familia in September, too. I really enjoyed it.”

HearsayExp is different from OwnExp from the viewpoint of how the experience is conveyed. Here, the system describes the experience as hearsay experience. For example, “I learned that the place is enjoyable because I read some internet reviews that described it.” In this study, we do not use such phrases as “I heard the place is enjoyable from my friend who told me that she enjoyed Sagrada Familia,” though these phrases are also hearsay experience. The reason is that the expression suggests that this robot has a relationship with others. It has been reported that people relate to a robot as more human-like when they know that the robot has a relationship with others [Cordar *et al.*, 2014]. Therefore, we decided not to use these phrases to isolate this effect.

We also developed two more systems to evaluate the effects of mentioning a reason for the impression. To directly evaluate the effects of adding such reasons, we used the following two methods that combined subjective and hearsay experiences.

Experience chooses a system experience that is similar to the user’s experience and adds a fact about its visit, an impression, and a reason for this impression. For example, when the same utterance “I saw Sagrada Familia” is given, the system says, “Oh, I went to Sagrada Familia in September, too. I really enjoyed it” as OwnExp or “I’ve read internet reviews that described it as enjoyable” as HearsayExp.

Exp+Reason also chooses a system experience that is similar to the user’s experience and mentions a fact about its visit, its impression, and a reason for this impression. For example, when the same utterance “I saw Sagrada Familia” is given, the system says, “Oh, I visited Sagrada Familia, too. I really enjoyed it because that is such a great building” as OwnExp with an impression reason or “I’ve heard that that place is enjoyable because I read internet reviews that described it as a great building” as HearsayExp with an impression reason.

All the dialogues were conducted based on the same dialogue flow, and the empathic utterances are different in each method. The dialogue contents change based on the user’s

utterances. We compared the above chat systems to evaluate the effects of sharing experiences depending on the expression types.

Note that we used simple empathy utterances as a baseline instead of not using empathy utterances because the effectiveness of simple empathy utterances is reported in the conventional research [Higashinaka *et al.*, 2008]. Besides, the interest in our work is the difference between methods of expressing empathy.

4.2 Experience Dataset Preparation

To fit the utterance template described in Section 3, we developed the system’s experience dataset. We chose the *travel* topic as our corpus domain because it evokes ordinary conversation in human-human interaction [Arimoto *et al.*, 2019] and it includes many experiences. Our system’s experience dataset was created by employed workers to whom we gave the following instructions: “Fill out the [When], [Where], ... items based on your experience or imagination. These items will be used as the contents of such utterance templates as ‘I visited [where] before.’” We provided the workers with three kinds of utterance templates. Since we chose travel as our corpus domain, creating conversations about famous tourist spots or famous foods is simple for each area by referring to guidebooks. The individuals who made the experience corpus mined their own memories of spots and filled in the 5W1H + impression items and their reasons. If they had never visited a particular tourist spot, they filled in the items by imagining information based on the guidebooks or the web. We focused on imagining the experience of being able to instead of the experience of not being able to because we considered that it is possible to easily generate empathy utterances by converting the user’s utterance when the user cannot have the experience. Moreover, the experience of being able to is more realistic for conveying empathy than that of not being able to. We collected each data in Japanese for this experiment.

We narrowed the tourist spots, food, and activities to Japan because a huge amount of experience data would be required if the area included the entire world. However, it is possible to easily and automatically extend the dataset using the phrase extractor described in Section 3. The phrase extractor can extract the target phrases from some sentences to fill each slot of the 5W1H frames. Therefore, it is possible to automatically generate the data from blogs or articles. To apply the system to other domains, we need to decide on suitable frames of the experience dataset. However, the phrase extractor can be applied using a small dataset, and the experience

dataset can be extended using the phrase extractor. Though we use only the travel domain to evaluate the effects of showing experience to convey empathy in this study, we will apply it to other domains in our future work.

To make the system easily find experiences that resemble the user’s experiences, each data in the corpus includes area information. Figure 3 shows an example of the experience/knowledge data in the dataset. Each data source in the experience corpus contains all the area information, 5W1H + impression items, and their reasons. We collected 2,652 data as shown in Table 3 for our experience corpus.

Item	Contents
Area (Prefecture)	Tokyo
Area (City)	Akihabara
When	in this summer vacation
Where	electronics shop
Who	with friends
What	bought electronics parts
Impression	it was exciting
Impression Reason	there were various types of parts

Table 3: One of the examples in the prepared experience dataset.

For the 5W1H + impression recognizer, we automatically generated a training dataset by using the collected experience corpus with sentence templates. We created eight possible template patterns as sentences such as “I [what] [who]” and “I [what] [impression],” and then filled each item into the sentence templates by changing the experience data in the experience corpus. The BIO tags were given at the same time as shown in Section 3.4. The generated training dataset is 213,204 sentences with BIO tags. We adopted BERT for sequence-labeling and a pretrained model trained with Japanese Wikipedia [Kikuta, 2019]. We then fine-tuned the pretrained BERT using our 213,204 dataset for sequence labeling.

In preliminary validation testing, the detection accuracy for the 5W1H items using the developed test dataset (number of datasets is 13,893) was 72%. Using this recognizer, the system can extract words or phrases for the 5W1H with impressions. We also confirmed that new types of phrases could be extracted as targets: “the park near Kyoto Station” is extracted as a location, even without a formal proper name, and “ate sushi” is extracted as a *what* item. The results extracted by the conventional-named entity recognizer and those by our proposed phrase recognizer are shown in Table 4. With the recognizer, our developed systems understood the context by filling in the 5W1H with impression frames through conversations.

If the park near Kyoto station is recognized as a user’s experience, the system cannot find the exact matched location in the system’s experience/knowledge dataset. Therefore, the system chooses an experience about a park without a named entity and conveys empathy using the experience at the park in this other location as an experience that resembles the user’s.

4.3 Experiment Procedure

We evaluated our developed system through user evaluations. We hired crowdworkers as examinees who can speak Japanese at the everyday conversation level to evaluate the systems. They talked with the five systems described in the previous section using Telegram,¹ which is a text-chat messaging tool, and subjectively evaluated each with questionnaires.

The following is the experimental procedure for each examinee: (i) The examinee answered some questions about the user characteristics. We asked about gender, familiarity with chat systems, and expectation of chat systems, but did not use the results in this study. (ii) The examinee spoke with system 1 by text chat. (iii) The examinee evaluated the system by answering the questionnaire in Table 5. (iv) Procedures (ii) and (iii) were repeated for each chat system. Each chat consisted of 30 utterances, fifteen from both the system and the users. The system said thank you and ended the conversation after the user input 15 utterances to the system. The examinees talked with all five systems. To avoid any order effect, the order of the conversation systems was randomly determined.

The examinees were instructed as follows: *Talk with five different AI robots in text chat and evaluate each system, respectively. Start from “Hello” input. Use multiple utterances in a single input to take your turn. Please talk with the AI robots freely and as naturally as possible, as you would speak with humans.* We then showed a list of tourist spots or food in each area with which the system was familiar in case the user does not remember any memories about his/her travel. Although our examinees were given the list, they were allowed to mention places, food, or other information outside of the list because we prioritized having the examinees speak freely and naturally without any feelings of constraint.

We prepared the four questions listed in Table 5, which were chosen for the following reasons: First, Q1 (the system talks based on its own experience) is asked to verify whether the proposed system is designed as intended. Q2 (impression of understanding) evaluated the impressions of the system’s ability, and Q3 (impression of empathizing) evaluated the scores that directly indicate the specific effects related to the purpose of our study. Q4 is derived from questionnaires that are often used to evaluate chat dialogue systems. In the preliminary examination, we asked “How did you enjoy the chat conversation with the system?” and “Did you feel satisfied with the chat conversation with the system?” and found that they were highly correlated to Q4. Therefore, we used Q4 to evaluate the secondary effects of talking about the system’s experiences. Each was evaluated using a 7-level Likert scale [Allen and Seaman, 2007]. Score 1 is lowest and score 7 is highest. To evaluate the significant differences among the scores, we calculated the *p* value using Welch’s t-test.

We hired 58 examinees (29 men and 29 women) ranging in age from 20s to 50s and had them evaluate the five systems. We conducted a preliminary experiment using another chatting task and selected 58 examinees who worked diligently without cutting corners for the purpose of screening. They repeated chatting with a system and answered the questionnaire

¹<https://telegram.org>

User utterance (<i>Italic: location phrase</i>)	NE extractor	Proposed phrase extractor
I went to <i>Italy</i> for summer vacation.	Italy	Italy
I saw a rainbow at <i>the park near Paris Station</i> .	Paris Station	the park near Paris Station
I often go to <i>electrical shops</i> .	N/A	electrical shops

Table 4: Comparison between location phrases extracted by the conventional method and by the proposed.

Qid	Question
Q1	How did you feel when the system talked based on its own experience?
Q2	How well do you think the system understood the dialogue context?
Q3	How well do you think the system empathized with you?
Q4	Would you like to talk with the system again?

Table 5: Questionnaires

five times. Since dialogue histories were submitted with each questionnaire, we collected 290 dialogue histories and the answers to the questionnaires.

Since we did not set any constraints to prioritize having examinees speak freely with the system, it is possible that some places, foods, or other information mentioned by the examinees fell outside of the system’s experience. In this situation, the system might confuse the mentioned place with a different but similar place that it knows. Hence, it was difficult to produce the implied empathic utterances because non-similar experiences were shared. This can also lead to dialogue breakdowns where the system and the examinee mention different places and create a situation in which the user feels that “the system does not understand me.”

Therefore, we analyzed the ratio of dialogues including dialogue breakdowns after the experiment. We hired an annotator to judge whether the conversation included dialogue breakdowns. The annotator was trained in advance to accurately annotate dialogue breakdown labels defined in the research [Higashinaka *et al.*, 2015], and she then judged whether the conversation included dialogue breakdowns or not by checking the dialogue logs. The conversation including dialogue breakdowns consisted of the following: The system talked about a tourist spot that was different from the one mentioned by the user because the system could not find any experience that resembled the user’s. Their conversations did not mesh because the system misunderstood the target that the user mentioned.

5 Evaluation Results

The questionnaire scores are shown in Figure 6. There is no significant difference in the results of (a) and (d). However, there are significant differences between OwnExp and HearsayExp in the results of (b) ($p=0.006$), that between OwnExp and HearsayExp ($p=0.01;0.05$). The results suggest that showing the reason of empathy using its own experience does not provide a rationale for empathy and may make the

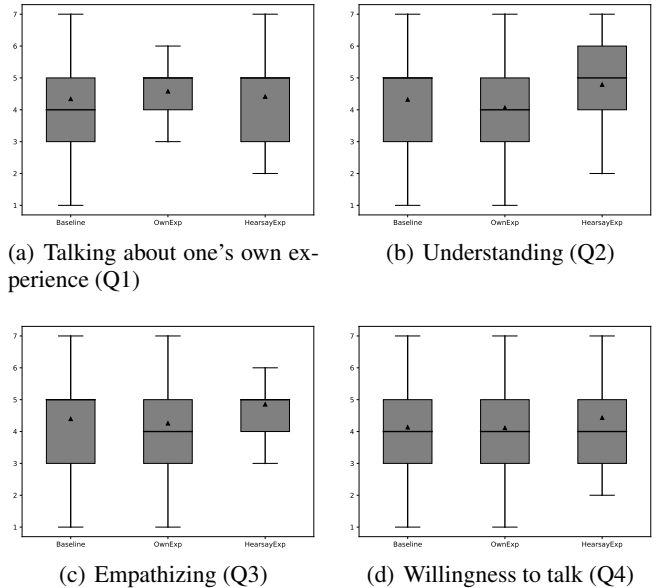


Figure 6: Effectiveness of talking experience compared with baseline. Each bar shows baseline, OwnExp, and HearsayExp results.

users feel it is disingenuous. On the other hand, showing the reason of empathy using the information obtained from its web surveys is more trustworthy than its own experience and can provide a rationale for empathy. The ratios of dialogues including dialogue breakdowns in each method are as follows: Baseline: 0.62, OwnExp: 0.65, and HearsayExp: 0.57. There is from a 0.03 to 0.08-point difference between the numbers of each. The difference is not so large compared to the above suggestion because the maximum difference is 5 (= 58 dialogues \times 0.08). However, what kinds of dialogue breakdowns are included in each dialogue may affect the understanding and empathy scores. We need to analyze the influence on the user’s impression by annotating the dialogue breakdown labels to each utterance in our future work.

We also evaluated the effectiveness of talking about the experience impression reasons (Figure 7). In the same manner, as with the previous results, there is no significant difference in the results of (a) and (d). However, there are significant differences between OwnExp and HearsayExp in the results of (b) ($p=0.0268<0.05$), that between OwnExp and HearsayExp ($p=0.031<0.05$). The results suggest that showing the reason of impression is effective to convey empathy. Moreover, the ratios of dialogues including dialogue breakdowns in each method are follows: Baseline: 0.62, Exp: 0.57, and Exp+Reason: 0.64. Although the ratio of dialogues includ-

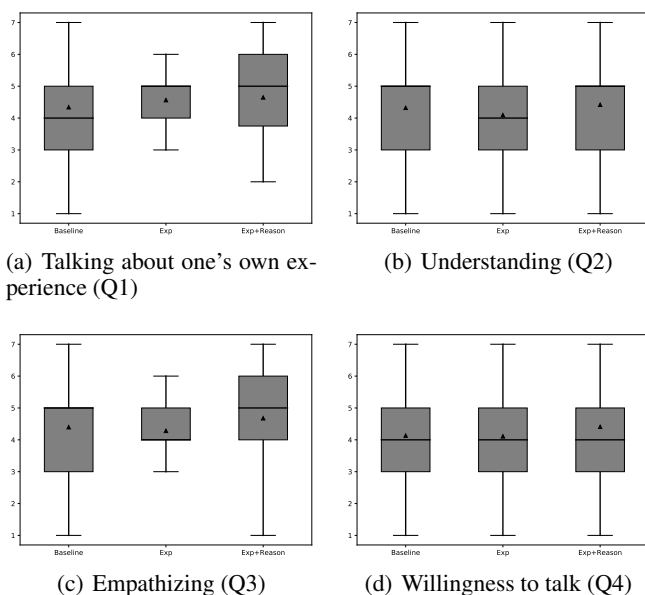


Figure 7: Effectiveness of talking about impression reason of experience compared with baseline. Each bar shows baseline, Exp without Reason, and Exp+Reason results.

ing dialogue breakdowns in Exp+Reason is highest in all, the scores of understanding and empathizing are highest in all, thus confirming the effects of showing the reason of impression described above.

To summarize the results of Figures 6 and 7, showing experiences as knowledge such as hearsay experiences and showing the reason of impression are useful to improve the sense of receiving empathy.

Finally, we discuss the evaluation results by showing the example of dialogue history shown in Figure 8, which is an actual dialogue history translated in English with a user and a system with HearsayExp. The number of empathic utterances that occur in a dialogue is three. However, sometimes this number changes depending on whether the system finds experiences that resemble the user's. When the utterance is indicated in bold letters, the system cannot find a suitable experience and only provides a simple utterance. The impression of empathy can be further improved if the system finds a suitable experience every time. Therefore, automatically increasing the experience data in the experience corpus through real human-agent conversation or extracted from human-human conversation is our next challenging work. This work is challenging because it includes the difficulties of judging whether the information retrieved from the user's utterance is collected or not.

6 Conclusion

To show empathy based on a system's constructed experience in chat-oriented dialogue systems, we proposed a dialogue system that shares its experiences that resemble a user's experiences as evidence of empathy. Through our experiments, we found that a dialogue system based on sharing its experience as knowledge or a hearsay experience is an effective

approach for showing empathy. This strategy improved the user's impressions of the system's ability to empathize. Furthermore, adding impression reasons when the system speaks about its experience effectively improved the user's impression that "the system empathizes with me!," even if the dialogue included dialogue breakdowns.

Through the evaluation, we also found that improving the impression of being empathized with is not related to improving the score of "willingness to talk." Since the score (willingness to talk) is often used to evaluate chat-oriented dialogue systems, it is essential to select a suitable score to evaluate the impression of the systems.

In this study, we could not analyze whether the sorts of dialogue breakdowns included in the dialogues affect the users' impression. We will try to divide the dialogue history into types of system misunderstandings and evaluate the impression of being empathized and understood in detail.

Furthermore, we will work to automatically collect a corpus that can be used as the system's pseudo experience, since this study's corpus was made by hand on a trial basis. We will use the 5WH recognizer we developed to understand the user's experience to collect the system's experience or knowledge through conversation or the web.

References

- [Allen and Seaman, 2007] I Elaine Allen and Christopher A Seaman. Likert scales and data analyses. *Quality progress*, 40(7):64–65, 2007.
- [Arimoto *et al.*, 2019] Tsunehiro Arimoto, Hiroaki Sugiyama, Masahiro Mizukami, Hiromi Narimatsu, and Ryuichiro Higashinaka. Analysis of satisfaction and topics in repeated conversation through days. In *Proceedings of the 23rd workshop on the Semantics and Pragmatics of Dialogue (SemDial2019)*, 2019.
- [Batson *et al.*, 1996] C Daniel Batson, Susie C Symptom, Jennifer L Hindman, Peter Decruz, R Matthew Todd, Joy L Weeks, Geoffrey Jennings, and Christopher T Burns. "I've been there, too": Effect on empathy of prior experience with a need. *Personality and social psychology bulletin*, 22(5):474–482, 1996.
- [Cheepen, 1988] Christine Cheepen. *The predictability of informal conversation*. Pinter Pub Ltd, 1988.
- [Cordar *et al.*, 2014] Andrew Cordar, Michael Borish, Adriana Foster, and Benjamin Lok. Building virtual humans with back stories: training interpersonal communication skills in medical students. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA 2014)*, pages 144–153. Springer, 2014.
- [DeVault *et al.*, 2014] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

U1: Hello.
 S1: Hello. I am considering taking a trip. Do you travel often? (Question)
 U2: Yes. I do.
 S2: Good. Where is a good place? (Simple utterance, follow-up question)
 U3: Beppu was good.
 S3: Beppu! Where did you visit in Beppu? (Repetition, follow-up question)
 U4: I joined the “Jigoku Meguri” tour.
 S4: Oh, Jigoku Meguri. How was it? (Repetition, follow-up question)
 U5: Nice. There are many jigokus, so I enjoyed them.
 S5: I see. When did you visit Oita?
 (Simple utterance instead of empathic utterance because similar experience was not found.)
 U6: In autumn.
 S6: Autumn sounds good. How about going in the summer? (Example empathic utterance, follow-up question)
 U7: Summer is also a good time to visit.
 S7: Oh really? I’ll keep that in mind. Did you eat any delicious food there? (Simple utterance, follow-up question)
 U8: Bird dishes are famous.
 S8: I see. Steamed food is famous in Oita. Did you have any?
 U9: I saw some around the hot springs.
 S9: An internet review called it “delicious.” Well, who did you go with? (Empathy utterance, follow-up question)
 U10: My family.
 S10: A family trip, nice. How about going with friends?
 U11: That’s nice, too.
 S11: I see. I’ll keep that in mind. I heard that Kannawa Onsen (hot spring) is famous in Oita. Did you go there?
 U12: It is also a famous different hot spring.
 S12: Oh, you went to an onsen. Was it relaxing? (Repetition, follow-up question)
 U13: Yes, it was.
 S13: I see. I’ve been thinking about going there because I read on the internet about a person who visited seven spas and claimed to be healed, both mentally and physically. (Empathy utterance)
 Did you buy any souvenirs? (Follow-up question)
 U14: Yes, always.
 S14: Good. What is your recommendation? (Simple utterance, follow-up question)
 U15: Kabosu manjyu (a Japanese bean-paste bun flavored with Kabosu, which is a type of citrus fruit).
 S15: I’ve never heard of that. I’ll remember that for future travel. Thank you for talking with me. (Ending)

Figure 8: Dialogue history with user (U) and system (S) with HearsayExp. Jigoku Meguri, referred to in U4 and S4, is a famous hot spring in Oita prefecture. S5 is related to U3 because Beppu is located in Oita prefecture.

- [Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [Eklund *et al.*, 2009] Jakob Eklund, TERESIA ANDERSSON-STRÅBERG, and Eric M Hansen. “I’ve also experienced loss and fear”: Effects of prior similar experience on empathy. *Scandinavian journal of psychology*, 50(1):65–69, 2009.
- [Fung *et al.*, 2016] Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. Towards empathetic human-robot interactions. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 173–193. Springer, 2016.
- [Grice, 1975] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.
- [Han *et al.*, 2013] Sangdo Han, Kyusong Lee, Donghyeon Lee, and Gary Geunbae Lee. Counseling dialog system with 5w1h extraction. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013)*, pages 349–353, 2013.
- [Higashinaka *et al.*, 2008] Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. Effects of self-disclosure and empathy in human-computer dialogue. In *Proceedings of 2008 IEEE Spoken Language Technology Workshop (SLT)*, pages 109–112. IEEE, 2008.
- [Higashinaka *et al.*, 2015] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 87–95, 2015.
- [Kawahara, 2018] Tatsuya Kawahara. Spoken dialogue system for a human-like conversational robot erica. In *In the Ninth International Workshop Spoken Dialogue Systems (IWSDS)*, 2018.

- [Kikuta, 2019] Yohei Kikuta. BERT pretrained model trained on Japanese wikipedia articles. <https://github.com/yoheikikuta/bert-japanese>, 2019.
- [McCallum and Li, 2003] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (CoNLL)*, volume 4, pages 188–191. Association for Computational Linguistics, 2003.
- [Morbini *et al.*, 2012] Fabrizio Morbini, Eric Forbell, David DeVault, Kenji Sagae, David Traum, and Albert Rizzo. A mixed-initiative conversational dialogue system for healthcare. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012)*, pages 137–139, 2012.
- [Morris *et al.*, 2018] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, 20(6):e10148, 2018.
- [Nadeau and Sekine, 2007] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.
- [Narimatsu *et al.*, 2018] Hiromi Narimatsu, Hiroaki Sugiyama, and Masahiro Mizukami. Detecting location-indicating phrases in user utterances for chat-oriented dialogue systems. In *Proceedings of the Fourth Linguistic and Cognitive Approaches to Dialog Agents Workshop (LaCATODA)*, pages 8–13, 2018.
- [Oh *et al.*, 2020] Young Hoon Oh, Kyungjin Chung, and Da Young Ju. Differences in interactions with a conversational agent. *International journal of environmental research and public health*, 17(9)(3189), 2020.
- [Preece *et al.*, 2001] Jennifer Preece, Kambiz Ghazati, et al. Observations and explorations of empathy online. *The internet and health communication: Experience and expectations*, 10(9781452233277):n11, 2001.
- [Tahara *et al.*, 2019] Shunichi Tahara, Kazushi Ikeda, and Keiichiro Hoashi. Empathic dialogue system based on emotions extracted from tweets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI 2019)*, pages 52–56, New York, NY, USA, 2019. Association for Computing Machinery.
- [Tsang, 2013] Mazie Tsang. The importance of empathy? as i have studied and experienced it. *Hawai'i Journal of Medicine & Public Health*, 72(9 Suppl 4):79, 2013.
- [Wu *et al.*, 2016] Bowen Wu, Baoxun Wang, and Hui Xue. Ranking responses oriented to conversational relevance in chat-bots. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 652–662, 2016.
- [Yan *et al.*, 2016] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64, 2016.
- [Zhang *et al.*, 2018] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.
- [Zhang *et al.*, 2020] Yichi Zhang, Zhijian Ou, and Zhou Yu. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611, 2020.