Investigating the Effect of Conveying Understanding Results in Chat-Oriented Dialogue Systems

Koh Mitsuda Ryuichiro Higashinaka Junji Tomita
NTT Media Intelligence Laboratories, Nippon Telegraph and Telephone Corporation
{mitsuda.ko,higashinaka.ryuichiro,tomita.junji}@lab.ntt.co.jp

Abstract

In dialogue systems, conveying understanding results of user utterances is important because it enables users to feel understood by the system. However, it is not clear what types of understanding results should be conveyed to users; some utterances may be offensive and some may be too commonsensical. In this paper, we explored the effect of conveying understanding results of user utterances in a chat-oriented dialogue system by an experiment using human subjects. As a result, we found that only certain types of understanding results, such as those related to a user's permanent state, are effective to improve user satisfaction. This paper clarifies the types of understanding results that can be safely uttered by a system.

1 Introduction

Current dialogue systems often convey the understanding results of user utterances for confirmation and for showing understanding. Task-oriented dialogue systems repeat information provided by users by using understanding results of user utterances to confirm the content of user utterances (Litman and Silliman, 2004; Raux et al., 2005). Chat-oriented dialogue systems also need to confirm the content of user utterances and to show understanding so that the systems can be more affective.

However, some of the understanding results should not be conveyed to users. For instance, some utterances (e.g. "You are stubborn.") may be offensive and some (e.g. "It is summer.") may be too commonsensical. To create a dialogue system which conveys one's understanding results, we need to know what types of the results can be used as system utterances.

In this paper, focusing on chat-oriented dialogue systems, we investigate the effects of conveying understanding results of user utterances. Specifically, we investigate the types of results that can be conveyed to users without lowering user satisfaction. For this purpose, we first prepared various types of understanding results. Then, by a subjective experiment, we examined their individual effects on user satisfaction. Note that, in this paper, we focus on the effects of system utterances that convey understanding results "as they are"; that is, utterances are literally the same as understanding results.

As a result of the experiment, we found that user's temporary states during dialogue should not be conveyed and user's permanent states and information irrelevant to users themselves can be conveyed safely as system utterances. Our results are useful for creating a dialogue system that conveys understanding results.

2 Data of Understanding Results

For our investigation, we need to prepare understanding results categorized by their types. For this purpose, we use a corpus of *PerceivedInfo* collected in our previous work (Mitsuda et al., 2017). In this corpus, user utterances in chat-oriented dialogue are associated with the information that can be perceived/inferred by humans from these utterances. Such information is called *PerceivedInfo* (perceived information).

Figure 1 shows an example of a chat-oriented dialogue and their *PerceivedInfo* in the corpus. As stimuli for collecting *PerceivedInfo*, a Japanese chat-oriented dialogue corpus (Higashinaka et al., 2014) was used. *PerceivedInfo* was written by multiple annotators using natural sentences with

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 389–394,
Taipei, Taiwan, November 27 – December 1, 2017

Chat-oriented dialogue

U_i	***Speaker***: utterance
U_1	***A***: Hello, nice to meet you!
U_2	***B***: Nice to meet you too.
U_3	***A***: I feel autumn coming, how about you?
U_4	***B***: I think so too.
U_5	***B***: The cicadas have gotten quiet recently.
...	
U_{12}	***B***: Do you go anywhere interesting in autumn?
U_{13}	***B***: I'll visit Mt. Fuji if I feel up to it.
...	
U_{36}	***A***: Let's talk about this next time.
U_{37}	***B***: Okay.

Perceived information for U_{13}

B doesn't mind going a long way.
B drives a car.
B is active.
B likes going on pleasure trips.
B likes mountains.
B likes Mt. Fuji.
B likes autumn leaves around Mt. Fuji.
B likes the outdoors.
B lives in Kanto prefecture.
B lives near Mt. Fuji.
B would like ***A*** to be surprised.
Mt. Fuji is famous for autumn leaves.

Figure 1: Example of chat-oriented dialogue and perceived information in *PerceivedInfo* corpus. ***A*** and ***B*** correspond to speakers.

Level 1	Level 2	Level 3	Description	Example
Thought (55.1%)	Thought (35.8%)	Belief self (30.7%)	Speaker's belief for him/herself	***A*** likes watching TV.
		Belief other (5.1%)	Speaker's belief toward listener	***A*** agrees with ***B***.
	Desire (19.3%)	Desire (9.9%)	Speaker's desire relative to him/herself	***A*** wants to talk about Mt. Fuji.
		Request (9.4%)	Speaker's request to listener	***A*** wants to be praised by ***B***.
Fact (44.9%)	A's fact (37.9%)	Attribute (20.2%)	User-modeling information of speaker	***A*** is married., ***A*** can drink.
		Behavior (14.4%)	Speaker's action	***A*** drives a car., ***A*** is thinking.
		Circumstance (3.3%)	Background around speaker	***A*** is close with friends.
	Other fact (7.0%)	Certain fact (3.9%)	Certain fact irrelevant to speaker	Mt. Fuji is famous for red leaves.
		Uncertain fact (3.1%)	Uncertain fact irrelevant to speaker	A rice crop may fail.

Table 1: Classification of perceived information used for investigation. ***A*** and ***B*** correspond to speakers.

regard to each utterance in the dialogue.

Table 1 shows the classification of *PerceivedInfo* created in our previous work. These types were determined by manual clustering. The classification was evaluated by inter-annotator agreement among three annotators using 3,000 instances of *PerceivedInfo* with "Level 3" types. Fleiss' κ showed substantial agreement (0.69), indicating the validity of the classification.

In this work, we use the *PerceivedInfo* in this corpus as understanding results and investigate the effects of system utterances that convey *PerceivedInfo*. We also investigate how the effects are different depending on the types of *PerceivedInfo* in the classification.

3 Experiment

Using *PerceivedInfo*, we evaluated the effects of system utterances conveying the understanding results in an experiment. Below, we explain the procedure to create the utterances for the experiment and how we evaluate them.

Figure 2 shows the flow of preparation and evaluation. We first select pairs of *PerceivedInfo* and a user utterance used for writing that *PerceivedInfo* from the corpus. The writers rewrite or refer to the *PerceivedInfo* and utterance to create system utterances. Finally, raters evaluate them by questionnaire, giving a score to each utterance.

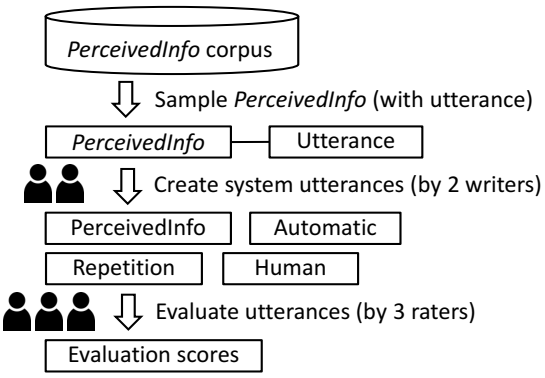

Figure 2: Preparation and evaluation of system utterances

3.1 Types of System Utterances

Table 2 shows the four types of system utterances prepared for evaluation. Utterances from *PerceivedInfo* are compared with those of three other types; namely, "Automatic," "Repetition," and "Human." "PerceivedInfo" is described below.

System utterance	Description	Example
PerceivedInfo	Confirmation of perceived information	You are active, aren't you?
Automatic	Response generated by a chat-oriented dialogue system	Do you work at Mt. Fuji?
Repetition	Repetition of user utterance in tag question form	You will visit Mt. Fuji, won't you?
Human	Response created by writer using keyword in user utterance	Mt. Fuji is a good place, isn't it?

Table 2: Types of system utterances prepared for evaluation. "Example" column shows system utterance when user utterance is "I'll visit Mt. Fuji if I feel up to it." (utterance U_{13} in Figure 1).

PerceivedInfo This utterance simply conveys *PerceivedInfo* in the form of confirmation. The utterance ends with a tag question form to confirm the content of *PerceivedInfo*. Rewriting *PerceivedInfo* is done manually. The symbols $\boldsymbol{A}$ and $\boldsymbol{B}$ that indicate speakers are changed to "You" or "I".

3.2 Types of Utterances for Comparison

We prepared three other types of system utterances for comparison. "Automatic" emulates the utterance of a chat-oriented dialogue system that is currently available. "Repetition" represents a simple repetition of the content of a user utterance. "Human" is an utterance conceived by human.

Automatic This utterance is an automatic response from a chat-oriented dialogue system that generates an utterance on the basis of keywords extracted from user utterances. To prepare responses, we use a Japanese chat-oriented dialogue system by NTT DOCOMO (Onishi and Yoshimura, 2014).

Repetition This utterance is a repetition of a predicate argument structure in a user utterance (Higashinaka et al., 2014). It ends with a tag question form (in Japanese, "*desu ne*") to show that the system understands the content of a user utterance. The utterance is manually created by extracting and rewriting a predicate argument structure from the user utterance.

Human This utterance is a human-level utterance (i.e., upper bound). We prepare it by having writers manually write an appropriate response to a keyword in the user utterance. Writers are instructed to select their favorite keyword in the utterance and use it to create a response that would satisfy users.

3.3 Preparation and Evaluation of Utterances

To clarify the difference of effects caused by the types of *PerceivedInfo*, we randomly selected approximately the same number of *PerceivedInfo* from each type in "Level 3" shown in Table 1. In total, we prepared 500 instances of *PerceivedInfo*; that is, 500 *PerceivedInfo* and user utterances associated with *PerceivedInfo*.

Using the 500 *PerceivedInfo* and utterances, "PerceivedInfo" and "Repetition" were written by a single writer and two versions of "Human" were written by two writers (Writer_1 and Writer_2) independently. We evaluated both utterances of "Human" written by the writers, because the quality of "Human" may depend on the writer. "Automatic" were generated from the chat-oriented dialogue system by NTT DOCOMO using the utterance as an input to the system. Following this experimental set-up, we prepared five types of utterances (including two versions of "Human") for each pair of *PerceivedInfo* and a user utterance, totalling 2,500 utterances, for evaluation.

To evaluate how each utterance is usable as a system utterance in dialogue, we annotated "naturalness" to the utterances. Raters were instructed to evaluate how natural the response was in the chat-oriented dialogue and to annotate an absolute score for each utterance in one of seven grades from one (very unnatural) to seven (very natural). They evaluated the five types of utterances at the same time. They could see not only a user utterance and system utterance but also the context before the user utterance. Three raters worked independently.

3.4 Results of Subjective Evaluation

Table 3 shows the results of the evaluation, where the average scores annotated by three raters to the five types of utterances are listed. The results show that "Human by Writer_1" and "Human by Writer_2" were ranked the highest by all raters, with "Automatic" ranked as the lowest. The order of the evaluated scores tended to be consistent in all raters ("Human by Writer_1," "Human by Writer_2," "Repetition," "PerceivedInfo," and "Automatic"). Spearman's rank correlation coefficient between two annotators averages at 0.56. "Per-

System utterance	Rater_1	Rater_2	Rater_3	Average
PerceivedInfo	2.7	3.1	3.1	3.0
Automatic	2.1	2.3	2.2	2.2
Repetition	3.7	4.0	4.6	4.1
Human by Writer_1	4.5	5.4	5.5	5.1
Human by Writer_2	4.5	5.1	5.5	5.0

Table 3: Naturalness scores of system utterances annotated by three raters

ceivedInfo" was evaluated as being more natural than "Automatic," but less natural than "Repetition."

From this result, we can say that using only *PerceivedInfo* as system utterances is not an effective method. However, since there may be a difference among the types of *PerceivedInfo*, we further investigated the evaluation scores in each type of *PerceivedInfo*.

Figure 3 shows the averaged naturalness scores annotated for each type of "PerceivedInfo." The scores were clearly divided into three ranges: 1–2, 2–4, and 4–5, and defined as **Low-rate type**, **Mid-rate type**, and **High-rate type**, respectively. We investigated what *PerceivedInfo* exist in each type and the reasons for their high or low rating. For reference, we list examples and scores of *PerceivedInfo* in each type in Table 4.

Low-rate type An utterance in the Low-rate type mainly refers to user's temporary states, such as thoughts, emotion, or behavior during dialogue (e.g., "You want me to agree, don't you?"). Even an utterance that includes a positive expression (e.g., "You like me, don't you?") tends to be evaluated as unnatural. This can be partly explained by the politeness theory (Brown and Levinson, 1987). Utterances in the Low-rate type that mention a user's temporary state would create the need for the user to explain. Thus, a user's negative face, the desire to be left free to act as he or she chooses, can be threatened.

Mid-rate type An utterance in the Mid-rate type generally refers to user's permanent states, such as favorites, experience, or profiles (e.g., "You like cool cars, don't you?"). Such an utterance tends to be evaluated as natural as "Repetition." However, an utterance including a negative expression (e.g., "You are stubborn, aren't you?") or a part of profiles (e.g. "You are a woman, aren't you?") tends to be evaluated as unnatural. This means that a mention of something negative or private about the user is not a good option. This can also be explained by the politeness theory as a violation of a user's positive face; that is a desire to keep self-image approved.

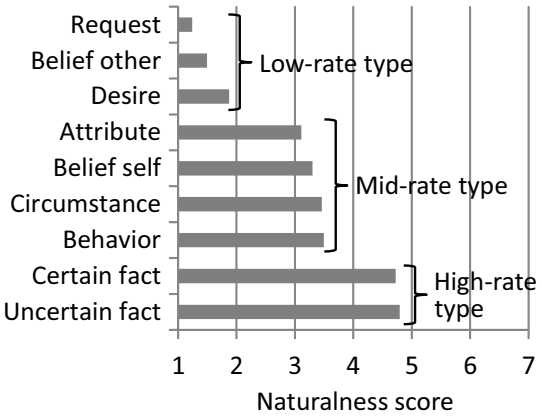

Figure 3: Naturalness scores of system utterances conveying perceived information on each type in Table 1

High-rate type An utterance in the High-rate type generally refers to the content that does not directly relate to users, such as general facts (e.g., "A trip abroad is expensive, isn't it?"). Many utterances are evaluated as more natural than "Repetition" and as natural as "Human." This may be because the utterances in the High-rate type do not threaten a user's face because the content of the utterances has no direct relation to users.

From the experiment, we found that utterances created from specific types of *PerceivedInfo* are evaluated as more natural than others. Our results conform to the politeness theory and further provide quantitative evaluation of utterances that convey *PerceivedInfo*. One interesting thing is that the violation of the negative face has more impact on the naturalness when compared to that of the positive face. It is of great interest that, although much *PerceivedInfo* occurs during a dialogue, only a part of it can be uttered.

Although further investigation is needed, our results are useful for providing a guideline for creating system utterances that convey understanding results.

Rate	PerceivedInfo	Worst utterance	Score	Best utterance	Score
Low	Request	You want me to agree, don't you?	1.0 (1,1,1)	You want to talk about China, don't you?	2.0 (1,2,3)
		You want to talk, don't you?	1.0 (1,1,1)	You want me to go to a gym, don't you?	1.7 (1,3,1)
		You want me to know about you, don't you?	1.0 (1,1,1)	You want to talk ordinarily, don't you?	1.7 (1,3,1)
	Belief other	You trust me, don't you?	1.0 (1,1,1)	You agree with me, don't you?	2.7 (3,3,2)
		You like me, don't you?	1.0 (1,1,1)	You are interested in my topic, aren't you?	2.3 (3,3,1)
		You misunderstand me, don't you?	1.0 (1,1,1)	My impression is changing, isn't it?	2.0 (3,2,1)
	Desire	You want to change the topic, don't you?	1.0 (1,1,1)	You long for nomadic life, don't you?	4.7 (5,2,7)
		You want to boast of your partner, don't you?	1.0 (1,1,1)	You want to go to Germany, don't you?	4.3 (4,4,5)
		You want to sympathize with me, don't you?	1.0 (1,1,1)	You want to live in cool place, don't you?	4.3 (4,5,4)
Mid	Attribute	You are a woman, aren't you?	1.0 (1,1,1)	You are sociable, aren't you?	6.3 (5,7,7)
		You are easygoing, aren't you?	1.0 (1,1,1)	You are willing to go out, aren't you?	6.3 (5,7,7)
		You are weak, aren't you?	1.0 (1,1,1)	You are kind, aren't you?	5.7 (3,7,7)
	Belief self	You are boastful, aren't you?	1.0 (1,1,1)	You like cool cars, don't you?	6.3 (5,7,7)
		You are a little embarrassed, aren't you?	1.3 (1,2,1)	You like Mt. Fuji, don't you?	6.0 (5,6,7)
		You are uninterested in agriculture, aren't you?	1.7 (1,3,1)	You like the outdoors, don't you?	5.3 (5,4,7)
	Circumstance	There is a computer in your home, isn't there?	1.7 (2,2,1)	You belong to the soccer team, don't you?	6.0 (5,6,7)
		You live on your husband's earnings, don't you?	1.7 (2,2,1)	It is sunny around you, isn't it?	5.3 (3,6,7)
		Your parents are well, aren't they?	2.0 (2,3,1)	Your relatives like celebrations, don't they?	5.0 (4,4,7)
	Behavior	You think about what to say, don't you?	1.3 (2,1,1)	You went out this summer, didn't you?	5.3 (3,6,7)
		You show me interests, don't you?	1.3 (1,2,1)	You lost your appetite due to how the food looks.	5.3 (4,5,7)
		You think what to say next,don't you?	1.3 (2,1,1)	You drink herb tee, don't you?	5.0 (2,7,6)
High	Certain fact	It is September, isn't it?	3.0 (3,5,1)	Wheels are expensive, aren't they?	6.7 (7,6,7)
		Bikes have various price ranges, don't they?	3.3 (3,3,4)	It is humid, isn't it?	6.3 (7,5,7)
		Road bikes and bicycles are different, aren't they?	3.7 (3,4,4)	Curved handlebars are special, aren't they?	6.3 (6,6,7)
	Uncertain fact	Using computers takes a lot of time, doesn't it?	2.0 (1,4,1)	A trip abroad is expensive, isn't they?	6.7 (7,6,7)
		That is a bad restaurant, isn't it?	3.3 (2,3,5)	Nagatomo showed great activity, didn't he?	6.3 (7,5,7)
		Muscle pain arises the next day, doesn't it?	3.7 (4,4,3)	Germany is safe, isn't it?	6.3 (6,5,7)

Table 4: Best and worst three utterances conveying perceived information on each type in Table 1. "Score" column shows average score and each score annotated by three raters.

4 Related Work

Although there has been no studies that explored the effect of utterances conveying system's understanding results to users, there have been several that have explored what linguistic behavior can be used or how to utter contents in dialogue systems from the viewpoints of social aspects (especially on the politeness theory).

For example, Gupta et al. constructed a task-oriented dialogue system in the cooking domain in which utterance generation is performed on the basis of the politeness theory (Gupta et al., 2007). Wang et al. estimated the politeness of each utterance in a task-oriented dialogue system by using various features, such as insults or criticisms (Wang et al., 2012). Danescu et al. constructed a corpus in which politeness is annotated in online community data and constructed a model for estimating politeness using linguistic features, such as gratitude expressions or positive and negative lexicons (Danescu-Niculescu-Mizil et al., 2013).

5 Conclusion

In this paper, we investigated what types of understanding results can be used as system utterances. Using the corpus of *PerceivedInfo* (perceived information), we manually created and evaluated the utterances that convey *PerceivedInfo*. We found that certain types of *PerceivedInfo*, especially those related to a user's permanent state and information irrelevant to users themselves, are usable.

For future work, we want to construct a dialogue system that conveys the understanding results in the way we proposed. For this purpose, we need to create an automatic estimator of *PerceivedInfo*. In this work, we used the understanding results as they were; however, we can create various system utterances from *PerceivedInfo*, and, in such a case, other types of *PerceivedInfo* can be effectively used. We want to further pursue how we can make use of *PerceivedInfo* in dialogue systems.

References

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proc. ACL*. pages 250–259.

Swati Gupta, Marilyn A. Walker, and Daniela M. Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *Proc. International Conference on Affective Computing and Intelligent Interaction*. pages 203–217.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki

Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open domain conversational system fully based on natural language processing. In *Proc. COLING*. pages 928–939.

Diane J Litman and Scott Silliman. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Proc. NAACL-HLT*. pages 5–8.

Koh Mitsuda, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2017. What information should a dialogue system understand?: Collection and analysis of perceived information in chat-oriented dialogue. In *Proc. IWSDS*.

Kanako Onishi and Takeshi Yoshimura. 2014. Casual conversation technology achieving natural dialog with computers. *NTT DOCOMO Technical Journal* 15(4):16–21.

Antonie Raux, Brian Langner, Dan Bohus, and Alan W Black Maxine Eskenazi. 2005. Let's Go Public! taking a spoken dialogue system to the real world. In *Proc. Interspeech*. pages 885–888.

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black, and Justine Cassell. 2012. "love ya, jerkface": Using sparse log-linear models to build positive and impolite relationships with teens. In *Proc. SIGDIAL*. pages 20–29.