What information should a dialogue system understand?: Collection and analysis of perceived information in chat-oriented dialogue

Koh Mitsuda, Ryuichiro Higashinaka, and Yoshihiro Matsuo

Abstract It is important for chat-oriented dialogue systems to be able to understand the various information from user utterances. However, no study has yet clarified the types of information that should be understood by such systems. With this purpose in mind, we collected and clustered information that humans perceive from each utterance (perceived information) in chat-oriented dialogue. We then clarified, i.e., categorized, the types of perceived information. The types were evaluated on the basis of inter-annotator agreement, which showed substantial agreement and demonstrated the validity of our categorization. To the best of our knowledge, this study is the first to clarify the types of information that a chat-oriented dialogue system should understand.

1 Introduction

Dialogue systems can use preselected knowledge about a specific task for understanding user utterances in task-oriented dialogue [2, 11], but this framework cannot be used in chat-oriented dialogue because it does not have (or at least seems not to have) a clear information structure.

Current chat-oriented dialogue systems interpret a user utterance by converting it into understanding results such as: keywords (approximating the focus or important information of a dialogue) [4], dialogue acts (for determining user intentions) [8], predicate argument structures (for understanding events) [4], emotions (for carrying out an action fitting a user's emotion) [7], and user attributes (for personalizing a response) [6]. Although such information is used as understanding results of dialogue systems, it is not clear whether these types of information are sufficient for the understanding of chat-oriented dialogue systems.

Koh Mitsuda, Ryuichiro Higashinaka, and Yoshihiro Matsuo
NTT Media Intelligence Laboratories, Nippon Telegraph and Telephone Corporation
E-mail: {mitsuda.ko, higashinaka.ryuichiro, matsuo.yoshihiro}@lab.ntt.co.jp

To investigate what sorts of information chat-oriented dialogue systems must understand, we focused on the information that humans perceive from each utterance in a dialogue. We call such information "perceived information." Tasks that relate to perceived information have been emerging recently; such as conversation entailment [13, 14], irony detection [9, 5], and document enrichment [15, 16]. However, they only focus on a specific type of perceived information and do not provide an overall picture or categorization of perceived information.

In this paper, we report on the data collection and analysis of perceived information. We collected many instances of perceived information written by multiple annotators. We then had other annotators manually cluster the same type of instances for analyzing what types of perceived information exist. To evaluate the clustering results, we tested the inter-annotator agreement in annotating the types of perceived information. Through this analysis, we clarified the kinds of information that dialogue systems should understand from utterances in chat-oriented dialogue.

2 Collection of perceived information

Here, we describe how we collected the perceived information in chat-oriented dialogue. First, we prepared dialogues for which the perceived information would be written down. Second, we collected perceived information by having multiple annotators write the perceived information for each utterance in the dialogues.

2.1 Preparation of dialogues

The choice of dialogues is important because they will be the source of the perceived information. For collecting general and various perceived information, we used a Japanese chat-oriented dialogue corpus collected by Higashinaka et al. [4], containing 3,680 dialogues between two people on various topics. We randomly selected 30 dialogues, which totaled 1,103 utterances.

2.2 Collection procedure

We collected perceived information (hereafter, we refer to it as *PerceivedInfo*). We defined *PerceivedInfo* as the information that humans can generally understand from an utterance in dialogue even though the information may not be explicit.

The procedure of data collection is two-fold, as illustrated in Figure 1. It is similar to the data-collection procedure of conversation entailment, where entailed information is collected from conversational data [13] and contradictory event pairs from propositional data [10].

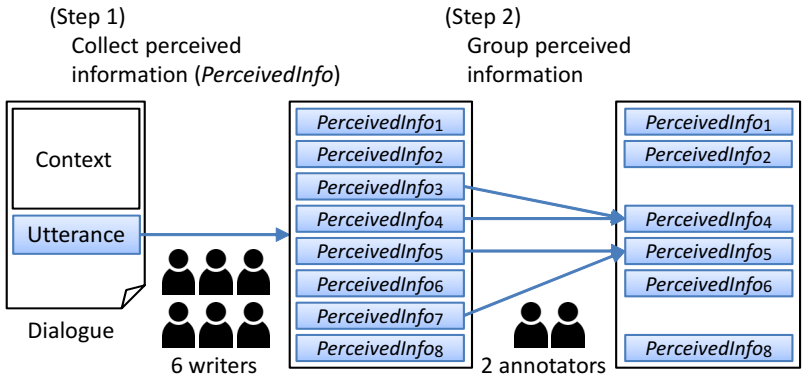

Fig. 1 Data collection and grouping of perceived information

(Step 1) First, annotators wrote *PerceivedInfo* as a natural sentence with regard to each utterance as a target in the dialogue. They wrote *PerceivedInfo* for all utterances in the dialogue in order from first to last. They could only use the context before the target utterance for writing *PerceivedInfo*; and the context after the target could not be used. They were instructed to write one or more *PerceivedInfo* for each utterance. We also told them that *PerceivedInfo* could not be a simple paraphrase of an utterance, complementation of omitted words, or information that is trivial on the basis of common sense, e.g., "We eat bread" or "Bread is made from flour." We made it mandatory that each *PerceivedInfo* had a predicate and an argument so that the proposition would be meaningful. We also made it mandatory that a *PerceivedInfo* should only be a single piece of information; thus, multiple pieces of information were divided into multiple *PerceivedInfo*. Six annotators worked independently in this step.

(Step 2) Second, duplicated or semantically similar *PerceivedInfo* for each utterance were grouped for the later process of clustering. Annotators who were not writers in Step 1 initially grouped *PerceivedInfo* independently and consulted one another to come up with the final results. If the content of multiple *PerceivedInfo*, such as content word, modality, tense, and negation were the same, they were grouped as the same *PerceivedInfo*. A representative *PerceivedInfo* written with the simplest wording was selected from each group. The representative *PerceivedInfo* was used in the next process of clustering.

We recruited 12 annotators for Step 1 and two for Step 2. In Step 1, six annotators worked on half of the target dialogues, and the remaining six worked for other half. For collecting perceived information by ordinary people, we employed non-experts in linguistics; thus, they had different backgrounds. Their ages ranged widely from in their twenties to in their fifties. The male-to-female ratio was about 1:1.

We collected 12,723 *PerceivedInfo* instances in Step 1. The instances were grouped into 11,533 (91%) instances in Step 2. We use the grouped 11,533 instances of *PerceivedInfo* in the next step of the analysis. Detailed information on the

	Dialogue	*PerceivedInfo*
Unique sentence	1,094	8,794
Unique word	1,596	3,740
Sentence	1,257	11,533
Word	10,856	116,413

Table 1 Amount of collected perceived information

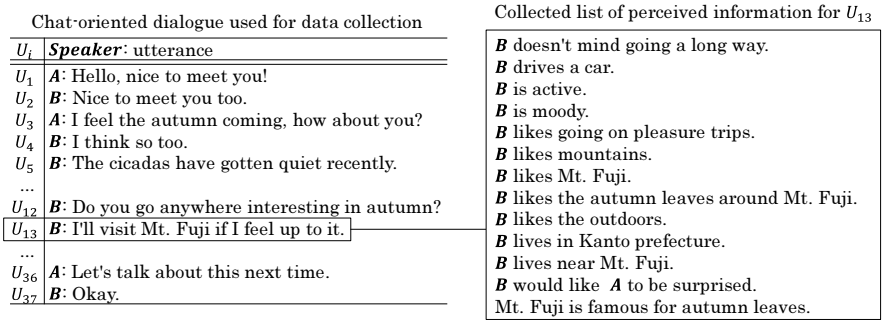

Fig. 2 Example of chat-oriented dialogue and perceived information for utterance

frequency of collected *PerceivedInfo* is shown in Table 1. We collected about ten instances of *PerceivedInfo* for each utterance, which suggests that various information can be perceived from an utterance. Figure 2 shows an example of a chat-oriented dialogue and *PerceivedInfo* collected for an utterance in the dialogue.

3 Clustering of perceived information

To investigate what types of information constitute *PerceivedInfo*, we clustered the collected *PerceivedInfo* by using multiple working groups. Note that the clustering was done manually by multiple annotators to ensure high-quality clustering.

3.1 Clustering procedure

The collected *PerceivedInfo* were clustered in two steps, as illustrated in Figure 3. First, multiple working groups made disjoint clusters of *PerceivedInfo*. Second, another working group hierarchically clustered all of the created clusters. The two steps are explained below:

(Step 1) Given instances of *PerceivedInfo*, multiple working groups independently and manually clustered similar *PerceivedInfo*. Two instances of *PerceivedInfo* were regarded as similar if both indicated the same type of information. We did not

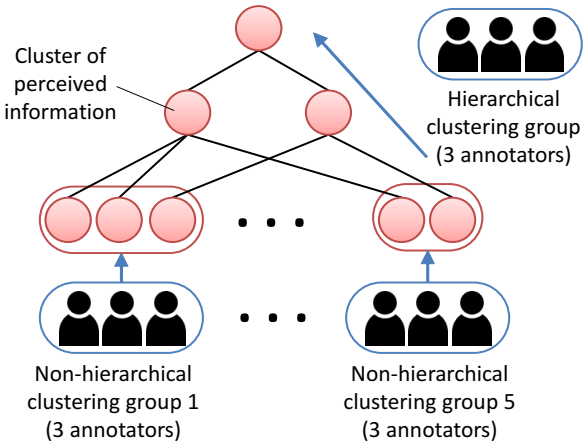

Fig. 3 Clustering procedure of perceived information

provide a rigid criterion of similarity; it is decided by each working group. They labeled each cluster indicating the type of *PerceivedInfo*. They continued clustering *PerceivedInfo* until all instances of *PerceivedInfo* were contained in some cluster.

(Step 2) Another working group merged the clusters created in Step 1. They manually organized the hierarchical clusters; that is, they found the most similar groups of clusters that had similar instances of *PerceivedInfo* and merged them into a new cluster. They repeated this process until there was only one cluster. As in Step 1, they labeled each cluster indicating the type of *PerceivedInfo* contained in the cluster.

We randomly selected 300 instances of *PerceivedInfo* from five dialogues and prepared 1,500 instances of *PerceivedInfo*. For Step 1, we assigned 300 instances of *PerceivedInfo* in each dialogue for each working group consisting of three annotators. For Step 2, another group consisting of three other annotators merged the clusters. We recruited 15 annotators for Step 1 and three annotators for Step 2. They were different from those who collected the *PerceivedInfo*. They had different backgrounds, and two experts in linguistics. The male-to-female ratio was about 1:1.

3.2 Types of perceived information

Table 2 shows the results of *PerceivedInfo* clustering; it is hierarchical on the basis of our clustering process. The right-most column, i.e., the fourth level, corresponds to the bottom-most clusters, and their integration progresses from right to left.

From Table 2, we can see that *PerceivedInfo* is divided into the speaker's "Thought" and "Fact." "Thought" consists of speaker's "Belief" and "Desire." "Fact" consists

First level	Second level	Third level	Fourth level
Thought (55.4%)	Belief (35.8%)	Belief self (30.7%)	Thinking (1.6%)
			The thing *A* is thinking (1.0%)
			Favorite (13.9%)
			Impression and evaluation (6.7%)
			Feeling (7.5%)
		Belief other (5.1%)	Impression of interlocutors (4.4%)
			Relation between *A* and *B* (0.7%)
	Desire (19.3%)	Desire (9.9%)	*A*'s desire related to *B* (3.1%)
			Self-contained desire (3.2%)
			A's desire 1: want to do (3.5%)
		Request (9.4%)	Wants interlocutor to do (1.1%)
			Request made to *B* (8.3%)
Fact (44.9%)	*A*'s fact (37.9%)	Attribute (20.2%)	*A*'s characteristics (19.5%)
			Possession (0.7%)
		Behavior (14.4%)	*A*'s past and experience (8.9%)
			A at present (5.5%)
		Circumstance (3.3%)	Circumstance of *A* and around *A* (1.6%)
			Circumstance around *A* (1.7%)
	Other fact (7.0%)	Certain fact (3.9%)	Fact about objects (0.7%)
			Objective fact 1: common things (0.3%)
			Objective fact (1.7%)
			Other fact: society (1.1%)
		Uncertain fact (3.1%)	Uncertain fact (1.2%)
			Fact (1.4%)
			Things that can happen (0.5%)

Table 2 Hierarchical clustering results of perceived information in chat-oriented dialogue. *A* corresponds to speaker, and *B* corresponds to listener (another speaker).

of facts regarding a speaker, namely "*A*'s fact," and "Other fact," which is information irrelevant to the speakers. These types on the second level are further divided into the types of the third level. The details of each type on the third level are given in Table 3. The clustering results show that *PerceivedInfo* has various types of information including ones used in conventional studies, such as the BDI model (Belief, Desire, and Intention) [12]. It is also clear that personal information is playing an important role in dialogue.

4 Evaluation

To evaluate the clusters, three annotators different from those who created the clusters annotated the labels of the *PerceivedInfo*. We used the first to third levels of clustering as annotation labels. Each annotator annotated 3,000 instances of *PerceivedInfo* that were not used in the clustering. They annotated labels by looking solely at *PerceivedInfo*, without using the context information that led to the *PerceivedInfo* in question.

Table 4 shows the inter-annotator agreement in terms of the label-wise agreement ratio and Fleiss' κ. κ values from 0.69 to 0.80 shows that there was substantial agreement between annotators. This quite high agreement indicates that the clusters cover various instances of *PerceivedInfo* and clearly distinguish each type of *PerceivedInfo*.

Belief self: Speaker's beliefs about his/herself
- Opinions: "*A* is displeased with prices in Tokyo." "*A* regards Japan as a safe country."
- Likes and Dislikes: "*A* likes playing TV games." "*A* hates smoking."
- Emotions: "*A* is excited." "*A* is happy at *B*'s praise."

Belief other: Speaker's belief toward the counterpart speaker
- Belief regarding utterances: "*A* agrees with *B*." "*A* can't believe *B*'s story."
- Belief regarding counterparts: "*A* is worried about *B*." "*A* thinks *B* is great."

Desire: Speaker's desire mainly relative to his/herself
- Desires of speakers: "*A* wants to go to Mt. Fuji." "*A* hopes summer ends soon."
- Desires relative to counterparts: "*A* wants to change the topic." "*A* wants to talk about his hobby."

Request: Speaker's requests to the counterpart
- Requests to counterparts: "*A* wants to be praised by *B*." "*A* wants to know about his hobby."
- Goals achieved with counterparts: "*A* wants to favor *B*'s opinion." "*A* wants for *B* to know what is interesting about the movie."

Attribute: User-modeling information of speakers
- Knowledge and Capability: "*A* knows a lot about cars." "*A* can drink."
- Social attributes: "*A* is woman." "*A* is married." "*A* lives in Kyoto."
- Personality: "*A* is earnest." "*A* is a determined person."

Behavior: Speaker's actions
- Habits: "*A* usually watches TV." "*A* hardly goes out." "*A* drives a car."
- Past and Experience: "*A* talked with his parents." "*A* has travelled abroad."
- State during dialogue: "*A* is trying to change the topic." "*A* seems proud." "*A* is thinking of what to say next."

Circumstance: Environment around speakers
- Relationships: "*A* is close with his parents." "*A*'s husband often watches TV."
- Living environment: "There are a lot of transfers in *A*'s job." "*A*'s parent's home is in Kanto prefecture."

Certain fact: Certain facts irrelevant to speakers
- Certain facts: "This summer is very humid." "Mt. Fuji is famous for autumn leaves."

Uncertain fact: Uncertain facts irrelevant to speakers
- Uncertain facts: "The rice crop may fail." "The economic depression is coming to an end."

Table 3 Descriptions and representative examples of perceived information in third level

Figure 4 shows the confusion matrices and numbers of labels as the annotation results. The confusion matrices indicated that every pair of annotators disagreed about the "Belief self" and "Attribute" types. A representative example of this disagreement is "*A* prefers natural food." The annotators also disagreed about the "Belief self" and "Behavior" types; sometimes, belief and behavior were difficult to separate as in the case of "*A* is blushing." Our brief analysis indicates the possible need to use a combination of labels in some cases.

The table on the bottom right shows the number of labels; the distributions of each annotator's results were mostly the same as in Table 2. This result suggests that the annotation is reliable and that different dialogues may share the same distribution of *PerceivedInfo*.

	First level	Second level	Third level
Label-wise agreement	0.90	0.86	0.75
Fleiss' κ	0.80	0.79	0.69

Table 4 Inter-annotator agreement as to types of perceived information from first level to third level in clusters of perceived information

a1 / a2	Belief self	Belief other	Desire	Request	Attribute	Behavior	Circumstance	Certain fact	Uncertain fact
Belief self	684	63	10	1	64	50	4	2	0
Belief other	13	130	1	1	3	10	0	0	0
Desire	37	9	252	72	1	13	0	2	0
Request	2	7	14	137	0	2	0	0	0
Attribute	29	3	0	0	466	39	16	1	0
Behavior	35	24	3	1	36	499	12	5	0
Circumstance	2	0	0	0	3	2	39	1	0
Certain fact	0	0	0	0	0	1	15	75	23
Uncertain fact	1	0	0	0	0	0	3	14	68

a1 / a3	Belief self	Belief other	Desire	Request	Attribute	Behavior	Circumstance	Certain fact	Uncertain fact
Belief self	710	68	13	0	44	36	2	3	2
Belief other	10	128	0	1	0	19	0	0	0
Desire	60	19	215	77	0	14	0	0	1
Request	2	17	18	123	0	2	0	0	0
Attribute	89	6	0	0	410	29	15	4	1
Behavior	64	35	1	0	75	419	15	5	1
Circumstance	3	0	0	0	2	1	40	1	0
Certain fact	2	0	0	0	0	2	11	79	20
Uncertain fact	3	0	0	0	0	0	2	40	41

a2 / a3	Belief self	Belief other	Desire	Request	Attribute	Behavior	Circumstance	Certain fact	Uncertain fact
Belief self	682	21	24	3	34	31	4	2	2
Belief other	9	194	5	2	1	25	0	0	0
Desire	45	12	166	52	0	4	0	0	1
Request	2	18	48	144	0	0	0	0	0
Attribute	115	4	0	0	426	19	4	4	1
Behavior	76	23	4	0	60	432	14	6	1
Circumstance	9	0	0	0	8	6	60	6	0
Certain fact	2	1	0	0	2	5	3	71	16
Uncertain fact	3	0	0	0	0	0	0	43	45

	a1		a2		a3		total	
	Freq	Ratio	Freq	Ratio	Freq	Ratio	Freq	Ratio
Belief self	878	0.29	803	0.27	943	0.31	2624	0.29
Belief other	158	0.05	236	0.08	273	0.09	667	0.07
Desire	386	0.13	280	0.09	247	0.08	913	0.10
Request	162	0.05	212	0.07	201	0.07	575	0.06
Attribute	554	0.18	573	0.19	531	0.18	1658	0.18
Behavior	615	0.20	616	0.21	522	0.17	1753	0.19
Circumstance	47	0.02	89	0.03	85	0.03	221	0.02
Certain fact	114	0.04	100	0.03	132	0.04	346	0.04
Uncertain fact	86	0.03	91	0.03	66	0.02	243	0.03
total	3000	1.00	3000	1.00	3000	1.00	9000	1.00

Fig. 4 Confusion matrices and number of labels used to annotate third-level types of perceived information. Symbols from "a1" to "a3" denote each annotator.

5 Conclusion

We investigated the types of information that humans understand in chat-oriented dialogue. To reveal what types of information constitute the perceived information, we collected a large amount of perceived information in chat-oriented dialogue and clustered it. The types of perceived information were evaluated on the basis of inter-annotator agreement, and their validity was verified. To the best of our knowledge, this study is the first to clarify the types of information that a system would require to understand in chat-oriented dialogue systems.

In the future, we intend to develop methods for automatically extracting perceived information from dialogue and build dialogue agents that can understand users better and take appropriate actions based on the estimated perceived information. As we now have categorical perceived information, we believe that we can initiate work on estimating perceived information. Another interesting future work

will be to discuss our results in terms of conventional dialogue theories, such as the cooperative principle [3], plan-based approaches to dialogue [2], and dialogue games [1].

References

1. Dialogue-games: Metacommunication structures for natural language interaction. Cognitive science **1**(4), 395–420 (1977)
2. Allen, J.: Recognizing intentions from natural language utterances. MIT Press (1983)
3. Grice, H.P.: Logic and conversation. In: P. Cole, J. Morgan (eds.) Syntax and semantics, vol. 3: Speech acts, pp. 41–58. Academic Press (1975)
4. Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y.: Towards an open domain conversational system fully based on natural language processing. In: Proc. COLING, pp. 928–939 (2014)
5. Joshi, A., Sharma, V., Battacharyya, P.: Harnessing context incongruity for sarcasm detection. In: Proc. ACL/IJCNLP, pp. 757–762 (2015)
6. Kim, Y., Bang, J., Choi, J., Ryu, S., Koo, S., Lee, G.G.: User information extraction for personalized dialogue systems. In: Proc. Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (2014)
7. Ma, C., Prendinger, H., Ishizuka, M.: Emotion estimation and reasoning based on affective textual interaction. In: Proc. Affective Computing and Intelligent Interaction (2005)
8. Meguro, T., Minami, Y., Higashinaka, R., Dohsaka, K.: Learning to control listening-oriented dialogue using partially observable markov decision processe. ACM Transactions on Speech and Language Processing **10**(15), 1–15 (2014)
9. Riloff, E., Quadir, A., Surve, P., Silva, L.D., Gilbert, N., huang, R.: Sarcasm as contrast between positive sentiment and negative situation. In: Proc. EMNLP, pp. 704–714 (2013)
10. Takabatake, Y., Morita, H., Kawahara, D., Kurohashi, S., Higashinaka, R., Matsuo, Y.: Classification and acquisition of contradictory event pairs using crowdsourcing. In: Proc. the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pp. 99–107 (2015)
11. Williams, J.D., Raux, A., Ramachandran, D., Black, A.: The dialog state tracking challenge. In: Proc. SIGDIAL, 404-413 (2013)
12. Wong, W., Cavedon, L., Thangarajah, J., Padgham, L.: Flexible conversation management using a BDI agent approach. In: Proc. IVA, pp. 464–470 (2012)
13. Zhang, C., Chai, J.Y.: What do we know about conversation participants: Experiments on conversation entailment. In: Proc. SIGDIAL, pp. 206–215 (2009)
14. Zhang, C., Chai, J.Y.: Towards conversation entailment: An empirical investigation. In: Proc. EMNLP, pp. 756–766 (2010)
15. Zhang, M., Qin, B., Liu, T., Zheng, M.: Triple based background knowledge ranking for document enrichment. In: Proc. COLING, pp. 917–927 (2014)
16. Zhang, M., Qin, B., Zheng, M., Hirst, G., Liu, T.: Encoding distributional semantics into triple-based knowledge ranking for document enrichment. In: Proc. ACL/IJCNLP, pp. 524–533 (2015)