Improving Taxonomy of Errors in Chat-oriented Dialogue Systems

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, Masahiro Mizukami

Abstract In previous studies, top-down and bottom-up approaches have been proposed for creating taxonomies of errors in chat-oriented dialogue systems. However, the reported κ (kappa) value for the taxonomy based on the top-down approach is low at 0.239, and no evaluation has been conducted for that based on the bottom-up approach. In this paper, we propose to revise these taxonomies to achieve better inter-annotator agreement. The revised taxonomy based on the bottom-up approach yielded a reasonable κ of 0.44 (Fleiss' κ), suggesting that this taxonomy can be used reliably to classify errors in chat-oriented dialogue systems.

1 Introduction

Current chat-oriented dialogue systems make various errors due to the limitation of current language technology [18, 1, 9, 17]. To improve such systems, error analysis is crucial; we need to have a good taxonomy of errors so that developers and researchers can have a common conception of errors, grasp the distribution of errors, and identify the main types of errors.

To the best of our knowledge, two taxonomies have been proposed for chat-oriented dialogue systems. One is based on the top-down (TD) (theory-driven) approach [7] and the other on the bottom-up (BU) (data-driven) one [10]. The former

Ryuichiro Higashinaka
NTT Media Intelligence Laboratories, NTT Corporation e-mail: higashinaka.ryuichiro@lab.ntt.co.jp

Masahiro Araki
Faculty of Information and Human Sciences, Kyoto Institute of Technology e-mail: araki@kit.ac.jp

Hiroshi Tsukahara
Research and Development Group, Denso IT Laboratory, Inc. e-mail: htsukahara@d-itlab.co.jp

Masahiro Mizukami
NTT Communication Science Laboratories, NTT Corporation e-mail: mizukami.masahiro@lab.ntt.co.jp

	Utterance (translated from Japanese)	NB	PB	B	TD taxonomy	BU taxonomy
U	Though watermelons are good, peaches are also yummy.					
S	Watermelons absent?? Peaches are good.	4	14	12	Utterance-Syntactic error	Grammatical error
U	(I) love sweet fruits.					
S	(I) like hard ones.	9	11	10	Response-Excess/lack of information	Lack of information
U	Talking of hard ones, I would say walnuts.					
S	(I) like Kazehaya.	6	5	19	Response-No-relevance	Diversion
U	(I) went to see the fireworks there.					
S	Kazehaya is so cool, isn't he?	2	5	23	Context-Non-relevant topic	Topic-change error
U	(He is) very cool.					
S	Thank you.	12	13	5	Context-Others	Others
U	Not at all.					
S	Yes.	16	6	8		
U	(I) almost got sick thanks to these very hot summer days.					
S	Will very hot summer days continue? The temperature will rise until midsummer, won't it?	9	14	7	Response-Unclear intention	Unclear relation
U	(I) can't sleep well on midsummer nights.					
S	Midsummer days are cool.	2	15	13	Environment-Lack of common sense	Violation of common sense
U	I don't like the summer heat.					
S	Watermelons are yummy on midsummer days.	11	13	6	Context-Excess/lack of proposition	Repetition

Fig. 1 Dialogue excerpt with error annotations by using revised TD/BU taxonomies. NB (Not a Breakdown), PB (Possible Breakdown), and B (Breakdown) indicate number of breakdown labels given to each system utterance. The dialogue system used here is NTT DOCOMO's chat API called DCM (See Section 4.1).

was derived by researchers' analyses of errors by a chat-oriented dialogue system. This taxonomy classifies errors by how each Grice's maxim is violated by a system utterance. The latter was derived by clustering annotators' comments that describe errors by a chat-oriented dialogue system. One way to evaluate the validity of a taxonomy is by using inter-annotator agreement; the reported κ (kappa) value for the TD taxonomy is low at 0.239, and no evaluation has been conducted for the BU taxonomy, suggesting that these taxonomies, in their current form, cannot be reliably used for error analysis.

We propose to revise these taxonomies to achieve better inter-annotator agreement. Our revised taxonomy based on the BU approach achieved a reasonable κ of 0.44 (Fleiss' κ). Figure 1 shows a dialogue example with error annotations by using our revised TD/BU taxonomies.

In the following section, we cover related work on the classification of errors in dialogue systems. In Section 3, we describe two current taxonomies of errors in chat-oriented dialogue systems. In Section 4, we describe how we revised the previously proposed taxonomies and report on their inter-annotator agreement. We also analyze the revised taxonomies on the basis of confusion matrices. Finally, in Section 5, we summarize the paper and mention future work.

2 Related work

There are mainly three approaches to classifying errors in dialogue systems.

One is to adopt the general taxonomy of miscommunication proposed by Clark [4]. According to Clark, there are four levels in communication; channel, signal, intention, and conversation. By using these four levels, errors can be classified into four categories depending on which level the errors occurred (e.g., [14], [3], [12]).

The second is to classify errors in view of cooperativeness in dialogue. Grice proposed a set of maxims of cooperative dialogue; quantity, quality, relevance, and manner [6]. [5] and [2] incorporated Grice's maxims in classifying errors in task-oriented dialogue systems. [7] used Grice's maxims to construct a taxonomy of errors in chat-oriented dialogue systems, although the inter-annotator agreement of this taxonomy was low, as we noted.

The third approach uses a clustering method to discover classes of errors. [10] showed the effectiveness of using a clustering method called the Chinese restaurant process (CRP) in clustering comments that describe errors by a system. In this study, a taxonomy of errors was automatically derived, although its validity was not evaluated.

In this paper, we verify the taxonomies derived from the second and third approaches. We did not use the first approach, four levels by Clark, because we currently deal with text-based systems in which channel and signal level errors rarely occur.

3 Current Taxonomies

We briefly describe the two current taxonomies of errors in chat-oriented dialogue systems.

3.1 Top-Down Taxonomy

The TD taxonomy [7] has main categories that distinguish to which scope of the context the errors relate; utterance-level, response-level (adjacency pair), context-level (local context), and environment-level (not within the local context) errors. In addition, within each main category, there are sub-categories based on Grice's maxims. The taxonomy has the following categories.

- Utterance-level (syntactic error, semantic error, un-interpretable)
- Response-level (excess/lack of information, non-understanding, no-relevance, unclear intention, misunderstanding),
- Context-level (excess/lack of proposition, contradiction, non-relevant topic, unclear relation, topic switch error)
- Environment-level (lack of common ground, lack of common sense, lack of sociality)

3.2 Bottom-Up Taxonomy

The BU taxonomy [10] has 17 categories. These categories were obtained from an analysis of comments (written by researchers) that describe the errors made by a chat-oriented dialogue system. Over 1,500 comments were collected and clustered using an automatic clustering method. As a result, 17 clusters were identified, which led to the following 17 categories:

> General quality, Not understandable, Ignore user utterance, Ignore user question, Unclear intention, Contradiction, Analysis failure, Inappropriate answer, Repetition, Grammatical

error, Expression error, Topic-change error, Violation of common sense, Word usage error, Diversion, Mismatch in conversation, and Social error

4 Revising Taxonomies

We evaluated the TD and BU taxonomies on the basis of inter-annotator agreement (we used agreement rate, Cohen's *kappa*, and Fleiss' *kappa*) and revised them iteratively. For the evaluation, we used datasets of multiple dialogue systems, which we believed necessary to ensure the generality of the resulting taxonomies. Note that the TD taxonomy has been evaluated, but only by using a dataset of a single system.

4.1 Datasets

We used the datasets provided by the two series of dialogue breakdown detection challenges (DBDCs) [8], i.e., DBDC and DBDC2.

DBDC dataset: The DBDC dataset provides 100 dialogues between human users and a chat-oriented dialogue system based on NTT DOCOMO's chat API [13] (called DCM). Each dialogue contains 21 utterances in which the first utterance is the system's prompt and the remaining are ten pairs of utterances in exchanges by a human user and the system. Dialogue breakdown labels (labels indicating whether a system utterance leads to a dialogue breakdown [11]) were given to each of the system's utterances by 30 annotators.

DBDC2 dataset: The DBDC2 dataset provides dialogues of three different chat-oriented dialogue systems. In addition to DCM, a system with label propagation [16] (called DIT) and a system based on statistical machine translation (called the IR-Status or IRS from [15]) were used. In the same manner as DBDC, 100 dialogues were collected for each system and the dialogue breakdown labels were given to each system utterance by 30 annotators; i.e., there were 300 dialogues with dialogue breakdown annotations.

We divided the data (400 dialogues in total) into five datasets (datasets A–E), each containing 80 dialogues (20 dialogues from the data of DCM in DBDC1 and those of DCM, DIT, and IRS in DBDC2). With these five datasets, we conducted five rounds of annotations and evaluations.

4.2 Evaluation of Current Taxonomies

We first evaluated the current taxonomies by using the first round of data (i.e., dataset A). We recruited four annotators and divided them up into two groups of two. One group used the TD taxonomy, the other the BU taxonomy. We provided them with annotation manuals that contained the definitions of the error categories with some examples. In this round, there were 16 category labels for the TD taxonomy and 17 for the BU taxonomy. Annotations were done to system utterances to which the majority (i.e. 15 or more) of the annotators of the DBDCs gave breakdown labels. This setting remained the same for all other rounds.

The inter-annotator agreement for the TD taxonomy was 0.35, and Cohen's κ was 0.26, which is rather low and similar to the previously reported κ [7]. When we focused on the main categories, that is, when we viewed the annotation as a four-label annotation, the agreement was 0.59, and Cohen's κ was 0.22, which was even lower than the sub-category annotation. The inter-annotator agreement of the BU taxonomy was 0.35, and Cohen's κ was 0.27, which is as low as that of the TD taxonomy.

4.3 Procedure for Revising Taxonomies

With κ values of 0.26 and 0.27 as a starting point, we iteratively revised the taxonomies with the following procedure.

1. Two annotators use the taxonomy to annotate one round of a dataset.
2. The inter-annotator agreement is calculated between the two annotators.
3. A confusion matrix is created, and on the basis of the discussion between the authors, the taxonomy is revised to reduce the confusion. In the revision, the annotation manual is also revised; the definitions are clarified and examples may be added. After the revision, go back to Step 1 for the next round.

We conducted five iterations in total. We used datasets B and C for the second and third rounds, respectively. We then re-used datasets A and B for the fourth and fifth rounds, respectively. We needed to re-use the data because the inter-annotator agreement did not improve as fast as we expected, and there was fear that the data would be exhausted before the inter-annotator agreement improved. Fortunately, we reached a certain level of agreement with the BU taxonomy (around 0.4 in Fleiss' κ) and decided to stop revising the taxonomies after the fifth round. The kappa for the TD taxonomy did not improve; it was around 0.2 throughout the rounds.

As the final step of the procedure, in the sixth round, we had crowd workers annotate the dialogues in dataset D with the revised taxonomies. The aim was to ensure that the errors can be reliably annotated by anyone, not just professional annotators. Below, we describe the revised TD/BU taxonomies and show the results of their evaluation.

4.4 Revised TD Taxonomy

The major improvements in the TD taxonomy are (1) revision of confusing categories and (2) introduction of a decision flow. The decision flow gives priority to error categories, making it possible to improve inter-annotator agreement in the case of possibly overlapping categories.

As a result of the confusion analysis of the trial rounds, we found that it is difficult to distinguish errors peculiar to human-system dialogue from other categories because it depends on one's ability to understand how a system works. Therefore, we merged such categories; we merged 'non-understanding' and 'misunderstanding' into 'non-understanding' in the utterance-level and 'non-relevant topic' and 'topic switch error' into 'non-relevant topic' in the context-level. In addition, we introduced an 'Others' category for each level as a natural consequence of introducing a decision flow. Our revised TD taxonomy is shown in Table 1.

In the TD taxonomy, the main category has to be determined first. The decision flow for the main category is shown in Fig. 2. The decision proceeds from narrower scope (utterance-level) to wider scope (context-level or environment-level).

Table 1 Revised TD taxonomy

Main category	Subcategory	Explanation
Utterance	Syntactic error	Grammatically invalid utterance
	Semantic error	Semantically invalid utterance
	Un-interpretable	Not understandable
	Others	Other utterance-level error
Response	Excess/lack of information	Utterance misses important information or contains unnecessary information
	Non-understanding	Content of utterance is false or inappropriate regarding previous user utterance
	No-relevance	Utterance does not have any relation to previous user utterance
	Unclear intention	Relation to previous user utterance is not clear
	Others	Other response-level error
Context	Excess/lack of proposition	Utterance is just empty words or repetition
	Contradiction	Utterance contains propositions that contradict what has been said
	Non-relevant topic	Topic of utterance is irrelevant to current context
	Unclear relation	Relation to previous dialogue context is not clear
	Others	Other context-level error
Environment	Lack of common ground	Utterance has no factual grounding
	Lack of common sense	Thoughtless utterance
	Lack of sociality	Offensive utterance
	Others	Other environment-level error

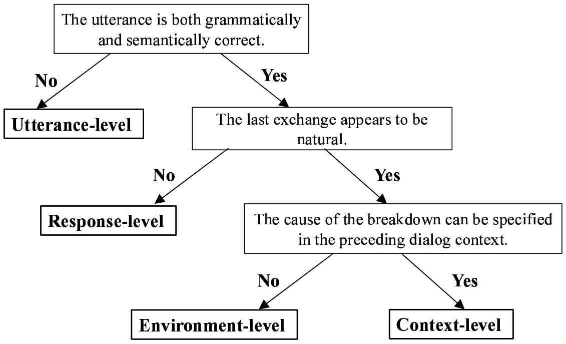

The utterance is both grammatically and semantically correct.
No
Yes
Utterance-level
The last exchange appears to be natural.
No
Yes
Response-level
The cause of the breakdown can be specified in the preceding dialog context.
No
Yes
Environment-level
Context-level

Fig. 2 Decision flow of main category for TD taxonomy

Once the main category is determined, the subcategory is decided following the decision flow for each level. In each decision flow of the subcategory, branching questions are ordered by their assumed ease of decision (Fig. 3).

4.5 Revised BU Taxonomy

The BU taxonomy also underwent (1) revision of confusing categories and (2) introduction of a decision flow.

We reduced the number of categories from 17 to 13. First, we removed ‘general quality’ and ‘mismatch in conversation’ because they were too vague and caused much confusion. We then merged ‘word usage error’ and ‘expression error’ because they focus on similar linguistic phenomena. We also merged ‘ignore user question’ and ‘inappropriate answer’ because it was found difficult to distinguish whether a question was ignored or answered inappropriately. In addition, we removed ‘unclear intention’ because the understanding of system intention greatly depended on the annotators.

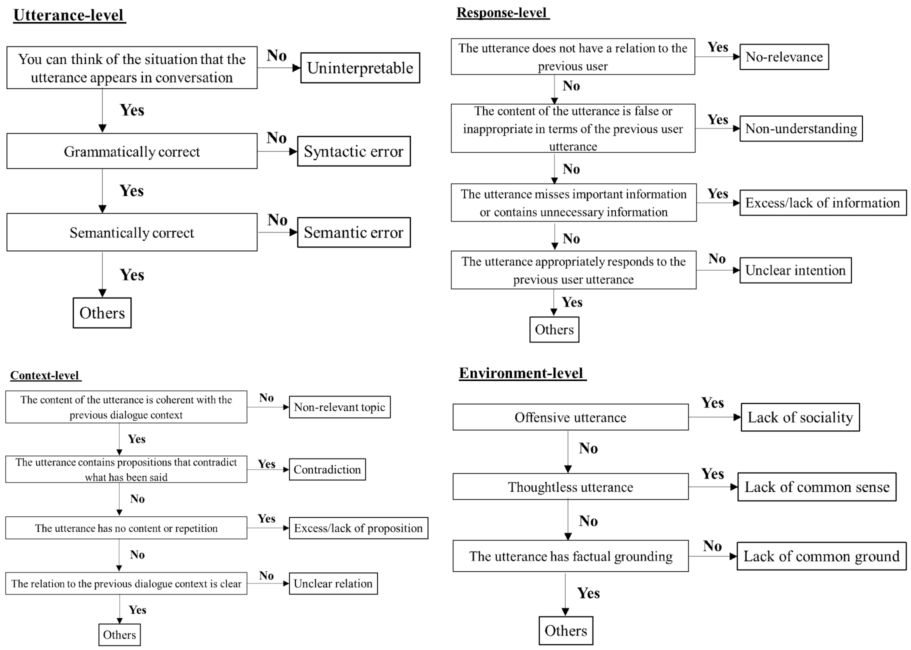

Utterance-level
You can think of the situation that the utterance appears in conversation
No
Uninterpretable
Yes
Grammatically correct
No
Syntactic error
Yes
Semantically correct
No
Semantic error
Yes
Others
Response-level
The utterance does not have a relation to the previous user
Yes
No-relevance
No
The content of the utterance is false or inappropriate in terms of the previous user utterance
Yes
Non-understanding
No
The utterance misses important information or contains unnecessary information
Yes
Excess/lack of information
No
The utterance appropriately responds to the previous user utterance
No
Unclear intention
Yes
Others
Context-level
The content of the utterance is coherent with the previous dialogue context
No
Non-relevant topic
Yes
The utterance contains propositions that contradict what has been said
Yes
Contradiction
No
The utterance has no content or repetition
Yes
Excess/lack of proposition
No
The relation to the previous dialogue context is clear
No
Unclear relation
Yes
Others
Environment-level
Offensive utterance
Yes
Lack of sociality
No
Thoughtless utterance
Yes
Lack of common sense
No
The utterance has factual grounding
No
Lack of common ground
Yes
Others

Fig. 3 Decision flow of subcategories for TD taxonomy

We initially removed ‘ignore user utterance’ because it was vague, but found afterwards the annotators had difficulty annotating inappropriate system responses to the user utterances that were not questions. Such errors were inevitably classified as ‘Others’. This actually leveraged the inter-annotator agreement, but we had many ‘Others’ errors, which is not desirable. Therefore, we borrowed two concepts from the TD taxonomy that relate to inappropriate responses, i.e., ‘lack of information’ and ‘unclear relation’. Table 2 shows the revised BU taxonomy, and Figure 4 shows the decision flow for the BU categories. As in the TD taxonomy, the order of the branching questions is by the assumed ease of decision.

4.6 Evaluation of Revised Taxonomies

As mentioned above, we ran a final round (sixth round) using crowd workers[1]. We used dataset D for the annotation and evaluation. For crowd-sourcing, we recruited two groups of ten Japanese-native workers for each taxonomy. The workers were in their 20’s to 50’s, and each group consisted of four females and six males. The workers were not experts in language-annotation tasks.

Table 3 shows the results. The inter-annotator agreement of the TD taxonomy was 0.32, Cohen’s κ was 0.24, and Fleiss’ κ was 0.21. We can say that the inter-annotator agreement is low. On the other hand, the inter-annotator agreement of the BU taxonomy was 0.54, Cohen’s κ was 0.44, and Fleiss’ κ was 0.44, which is

[1] `https://www.lancers.jp/`

Table 2 Revised BU taxonomy

	Category	Explanation
1	Not understandable	Utterance is un-interpretable.
2	Grammatical error	Utterance is ungrammatical.
3	Word usage error	Words that do not fit context are used in utterance.
4	Ignore user question	System ignores or fails to answer question.
5	Topic-change error	Utterance does not reflect topic introduced by user.
6	Diversion	Utterance abruptly introduces different topic.
7	Contradiction	Content of utterance contradicts what has already been said.
8	Repetition	Utterance is just repeated without new information.
9	Lack of information	Utterance misses important information.
10	Unclear relation	Relation between utterance and context is unclear.
11	Social error	Utterance lacks politeness.
12	Violation of common sense	Utterance lacks common sense.
13	Others	Miscellaneous error

Table 3 Inter-annotator agreement, Cohen's κ, and Fleiss' κ for our revised taxonomies when crowd workers were employed as annotators. Cohen's κ was derived by averaging the κ values of all pairwise combinations of crowd workers.

	Agreement	Cohen's κ	Fleiss' κ
TD	0.32	0.24	0.21
BU	0.54	0.44	0.44

reasonably high. We additionally conducted an annotation by using the professional annotators (four annotators) and found that the inter-annotator agreement between the professionals and crowd workers was sufficiently high (0.45 in Cohen's κ).

Figure 5 shows the agreement and κ values in each round. Despite our revision, the TD taxonomy did not improve much. However, we observed a steady improvement for the BU taxonomy. In the third round, the inter-annotator agreement was rather high. This was when many errors were labeled as 'Others' because of the lack of appropriate error categories. We successfully reached the same level of agreement by appropriately modifying the taxonomy.

Figures 6 and 7 show the confusion matrices for the TD and BU taxonomies in the form of heat maps. For the TD taxonomy, it seems that it was difficult to distinguish the scope of the response from that of the context because the response can be part of the context. In addition, it seems difficult to classify errors that occur after other errors because such a case is not considered in the dialogue theories (i.e., Grice's maxims) on which the TD taxonomy is based. In fact, we compared the inter-annotator agreement of the first breakdown utterance during dialogue (error after a normal situation) and that of the rest (error after errors). The Fleiss' κ was 0.32 for the former and 0.19 for the latter, suggesting that the TD taxonomy can be effective for the analysis of a system that involves fewer breakdowns.

For the BU taxonomy, there was a problem in distinguishing '9: lack of information' vs. '10: unclear relation'; it seems difficult to distinguish 'something is missing' vs. 'something is unclear'.

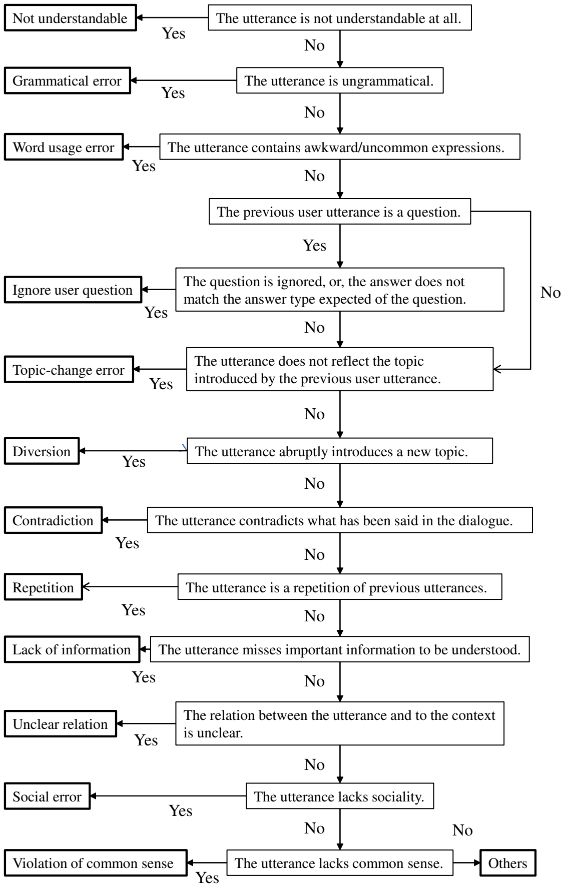

Not understandable
Yes
The utterance is not understandable at all.
No
Grammatical error
Yes
The utterance is ungrammatical.
No
Word usage error
Yes
The utterance contains awkward/uncommon expressions.
No
The previous user utterance is a question.
Yes
No
Ignore user question
Yes
The question is ignored, or, the answer does not match the answer type expected of the question.
No
Topic-change error
Yes
The utterance does not reflect the topic introduced by the previous user utterance.
No
Diversion
Yes
The utterance abruptly introduces a new topic.
No
Contradiction
Yes
The utterance contradicts what has been said in the dialogue.
No
Repetition
Yes
The utterance is a repetition of previous utterances.
No
Lack of information
Yes
The utterance misses important information to be understood.
No
Unclear relation
Yes
The relation between the utterance and to the context is unclear.
No
Social error
Yes
The utterance lacks sociality.
No
No
Violation of common sense
Yes
The utterance lacks common sense.
Others

Fig. 4 Decision flow for BU categories

5 Summary and Future Work

We revised two previously proposed taxonomies of errors in chat-oriented dialogue systems. The revised version of the BU taxonomy achieved reasonable inter-annotator agreement; 0.44 for both Cohen's κ and Fleiss' κ, making it possible for it to be safely used to classify errors in chat-oriented dialogue systems. These κ values were obtained by crowd workers (i.e., non-professional annotators), which is encouraging because it means that our taxonomy offers easy-to-understand conceptions of errors. We consider that the level of inter-annotator agreement that we achieved is reasonably high and can be used for classifying errors in chat-oriented dialogue systems.

For future work, we plan to further investigate the reason behind the poor inter-annotator agreement of the TD taxonomy. We noted in our analyses that it is difficult to distinguish between the response-level and context-level. One solution may be to

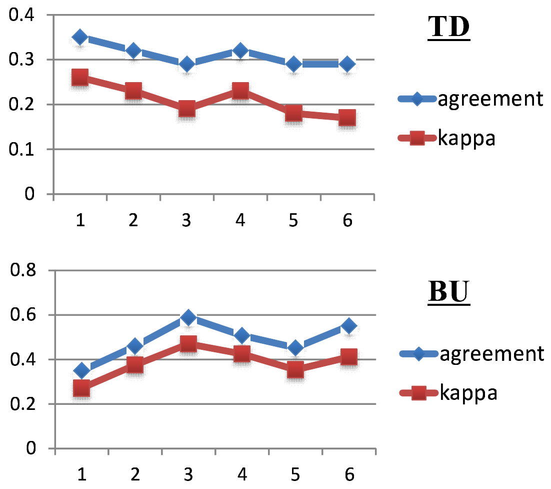
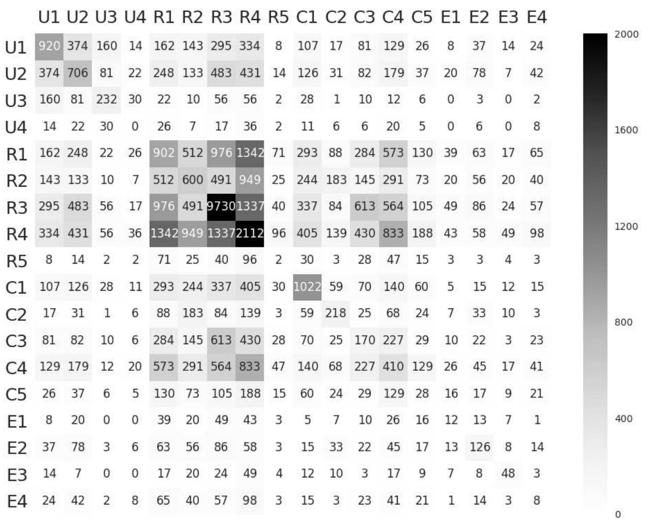


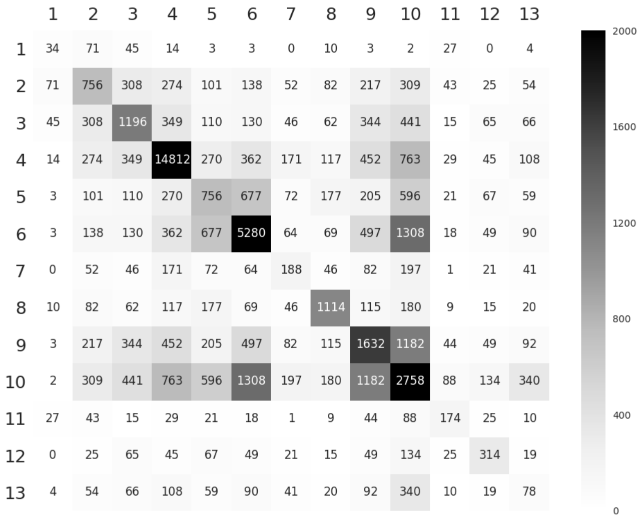
Fig. 5 Agreement and κ values in each round. We used professional annotators from rounds 1–5 and crowd workers in round 6.



	U1	U2	U3	U4	R1	R2	R3	R4	R5	C1	C2	C3	C4	C5	E1	E2	E3	E4
U1	920	374	160	14	162	143	295	334	8	107	17	81	129	26	8	37	14	24
U2	374	706	81	22	248	133	483	431	14	126	31	82	179	37	20	78	7	42
U3	160	81	232	30	22	10	56	56	2	28	1	10	12	6	0	3	0	2
U4	14	22	30	0	26	7	17	36	2	11	6	6	20	5	0	6	0	8
R1	162	248	22	26	902	512	976	1342	71	293	88	284	573	130	39	63	17	65
R2	143	133	10	7	512	600	491	949	25	244	183	145	291	73	20	56	20	40
R3	295	483	56	17	976	491	9730	1337	40	337	84	613	564	105	49	86	24	57
R4	334	431	56	36	1342	949	1337	2112	96	405	139	430	833	188	43	58	49	98
R5	8	14	2	2	71	25	40	96	2	30	3	28	47	15	3	3	4	3
C1	107	126	28	11	293	244	337	405	30	1022	59	70	140	60	5	15	12	15
C2	17	31	1	6	88	183	84	139	3	59	218	25	68	24	7	33	10	3
C3	81	82	10	6	284	145	613	430	28	70	25	170	227	29	10	22	3	23
C4	129	179	12	20	573	291	564	833	47	140	68	227	410	129	26	45	17	41
C5	26	37	6	5	130	73	105	188	15	60	24	29	129	28	16	17	9	21
E1	8	20	0	0	39	20	49	43	3	5	7	10	26	16	12	13	7	1
E2	37	78	3	6	63	56	86	58	3	15	33	22	45	17	13	126	8	14
E3	14	7	0	0	17	20	24	49	4	12	10	3	17	9	7	8	48	3
E4	24	42	2	8	65	40	57	98	3	15	3	23	41	21	1	14	3	8

Fig. 6 Confusion matrix for revised TD taxonomy. Prefixes U, R, C, and E stand for Utterance, Response, Context, and Environment main categories. U1–E4 mean error categories from Utterance-Syntactic error to Environment-Others.

merge these two levels. We also want to resolve cases in which one error category is the cause of the other. In our revision of the BU taxonomy, we borrowed two error categories from the TD, but it may be possible to merge BU and TD taxonomies for a better and complete taxonomy. Now that we have a reasonable taxonomy of



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	34	71	45	14	3	3	0	10	3	2	27	0	4
2	71	756	308	274	101	138	52	82	217	309	43	25	54
3	45	308	1196	349	110	130	46	62	344	441	15	65	66
4	14	274	349	14812	270	362	171	117	452	763	29	45	108
5	3	101	110	270	756	677	72	177	205	596	21	67	59
6	3	138	130	362	677	5280	64	69	497	1308	18	49	90
7	0	52	46	171	72	64	188	46	82	197	1	21	41
8	10	82	62	117	177	69	46	1114	115	180	9	15	20
9	3	217	344	452	205	497	82	115	1632	1182	44	49	92
10	2	309	441	763	596	1308	197	180	1182	2758	88	134	340
11	27	43	15	29	21	18	1	9	44	88	174	25	10
12	0	25	65	45	67	49	21	15	49	134	25	314	19
13	4	54	66	108	59	90	41	20	92	340	10	19	78

Fig. 7 Confusion matrix for revised BU taxonomy. 1–13 indicate error categories from Not understandable to Others (see Table 2).

errors, we also want to use this taxonomy for reducing errors of our chat-oriented dialogue systems.

References

1. Banchs, R.E., Li, H.: IRIS: a chat-oriented dialogue system based on the vector space model. In: Proc. the ACL 2012 System Demonstrations, pp. 37–42 (2012)
2. Bernsen, N.O., Dybkjær, H., Dybkjær, L.: Principles for the design of cooperative spoken human-machine dialogue. In: Proc. ICSLP, vol. 2, pp. 729–732 (1996)
3. Bohus, D., Rudnicky, A.I.: Sorry, i didn't catch that!–an investigation of non-understanding errors and recovery strategies. In: Proc. SIGDIAL, pp. 128–143 (2005)
4. Clark, H.H.: Using language. Cambridge university press (1996)
5. Dybkjær, L., Bernsen, N.O., Dybkjær, H.: Grice incorporated: cooperativity in spoken dialogue. In: Proc. COLING, vol. 1, pp. 328–333 (1996)
6. Grice, H.P.: Logic and conversation. In: P. Cole, J. Morgan (eds.) Syntax and Semantics 3: Speech Acts, pp. 41–58. New York: Academic Press (1975)
7. Higashinaka, R., Funakoshi, K., Araki, M., Tsukahara, H., Kobayashi, Y., Mizukami, M.: Towards taxonomy of errors in chat-oriented dialogue systems. In: Proc. SIGDIAL, pp. 87–95 (2015)
8. Higashinaka, R., Funakoshi, K., Kobayashi, Y., Inaba, M.: The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In: Proc. LREC, pp. 3146–3150 (2016)
9. 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)
10. Higashinaka, R., Mizukami, M., Funakoshi, K., Araki, M., Tsukahara, H., Kobayashi, Y.: Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In: Proc. EMNLP, pp. 2243–2248 (2015)

11. Martinovsky, B., Traum, D.: The error is the clue: Breakdown in human-machine interaction. In: Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems, pp. 11–16 (2003)
12. Möller, S., Engelbrecht, K.P., Oulasvirta, A.: Analysis of communication failures for spoken dialogue systems. In: Proc. INTERSPEECH, pp. 134–137 (2007)
13. Onishi, K., Yoshimura, T.: Casual conversation technology achieving natural dialog with computers. NTT DOCOMO Technical Journal **15**(4), 16–21 (2014)
14. Paek, T.: Toward a taxonomy of communication errors. In: Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems, pp. 53–58 (2003)
15. Ritter, A., Cherry, C., Dolan, W.B.: Data-driven response generation in social media. In: Proc. EMNLP, pp. 583–593. Association for Computational Linguistics (2011)
16. Tsukahara, H., Uchiumi, K.: System utterance generation by label propagation over association graph of words and utterance patterns for open-domain dialogue systems. In: Proc. PACLIC, pp. 323–331 (2015)
17. Vinyals, O., Le, Q.: A neural conversational model. arXiv preprint arXiv:1506.05869 (2015)
18. Wallace, R.S.: The anatomy of alice. In: Parsing the Turing Test, pp. 181–210. Springer (2009)