Integrated taxonomy of errors in chat-oriented dialogue systems

Ryuichiro Higashinaka[1]*, **Masahiro Araki**[2], **Hiroshi Tsukahara**[3], **Masahiro Mizukami**[4]
[1]NTT Media Intelligence Laboratories, NTT Corporation
[2]Faculty of Information and Human Sciences, Kyoto Institute of Technology
[3]Research and Development Group, Denso IT Laboratory, Inc.
[4]NTT Communication Science Laboratories, NTT Corporation
ryuichiro.higashinaka.tp@hco.ntt.co.jp, araki@kit.ac.jp
htsukahara@d-itlab.co.jp, masahiro.mizukami.df@hco.ntt.co.jp

Abstract

This paper proposes a taxonomy of errors in chat-oriented dialogue systems. Previously, two taxonomies were proposed; one is theory-driven and the other data-driven. The former suffers from the fact that dialogue theories for human conversation are often not appropriate for categorizing errors made by chat-oriented dialogue systems. The latter has limitations in that it can only cope with errors of systems for which we have data. This paper integrates these two taxonomies to create a comprehensive taxonomy of errors in chat-oriented dialogue systems. We found that, with our integrated taxonomy, errors can be reliably annotated with a higher Fleiss' kappa compared with the previously proposed taxonomies.

1 Introduction

From their social aspects, chat-oriented dialogue systems have been attracting much attention in recent years (Wallace, 2009; Banchs and Li, 2012; Higashinaka et al., 2014; Ram et al., 2018). Neural-based methods have been extensively studied and have yielded promising results (Vinyals and Le, 2015; Zhang et al., 2018; Dinan et al., 2019; Adiwardana et al., 2020; Roller et al., 2020). Yet, the performance of these systems is still unsatisfactory, causing dialogues to often break down.

One way to reduce the errors made by the systems is to understand what kinds of errors the systems are making and find solutions to counter them. For such a purpose, a taxonomy of errors will be useful. For task-oriented dialogue systems, several taxonomies have been proposed (Dybkjær et al., 1996; Bernsen et al., 1996; Aberdeen and Ferro, 2003; Dzikovska et al., 2009), leading to effective analyses for improving system performance. For dialogue systems that are chat-oriented, such taxonomies have also been proposed. Higashinaka et al. (2015a; 2015b) proposed two taxonomies; one is theory-driven and the other data-driven. However, the former suffers from the fact that dialogue theories for human conversation on which the taxonomy is based, such as Grice's maxims (Grice, 1975) and adjacency pairs (Schegloff and Sacks, 1973), are often not appropriate for categorizing errors made by chat-oriented dialogue systems. The latter has limitations in that it can only cope with errors for which we have data. Because of such shortcomings, these taxonomies suffer from low inter-annotator agreements, failing to successfully conceptualize the errors (Higashinaka et al., 2019).

This paper aims to create a new taxonomy of errors in chat-oriented dialogue systems. On the basis of the two taxonomies previously proposed, we discuss their merits and demerits, and we integrate the two into a comprehensive one. We verify the appropriateness of the integrated taxonomy by its inter-annotator agreement. We found that the kappa values were reasonable at 0.567 and 0.488 when expert annotators and crowd workers were used for annotation, respectively, and these values were much better than those of the previous taxonomies. This indicates that the errors have successfully been conceptualized, and we can safely use them to analyze errors made by chat-oriented dialogue systems.

2 Previous Taxonomies and Integration

Higashinaka et al. proposed two taxonomies of errors in chat-oriented dialogue systems: theory-driven (Higashinaka et al., 2015a) and data-driven (Higashinaka et al., 2015b).[1]

*Currently mainly affiliated with Nagoya University.

[1]Note that although Higashinaka et al. used "top-down" and "bottom-up" to name their taxonomies, we use "theory-driven" and "data-driven," which we consider to be more appropriate.

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 89–98
July 29–31, 2021.

The theory-driven taxonomy is based on principles in dialogue theories that explain the cooperative behavior in human dialogues. The taxonomy uses the deviations from such principles as error types. In contrast, the data-driven taxonomy uses the dialogue data of chat-oriented systems in order to identify typical errors made by such systems. The taxonomy was created by first collecting comments (textual descriptions) describing errors made by systems and then clustering the comments; each resulting cluster corresponds to an error type.

2.1 Theory-driven taxonomy

The theory-driven taxonomy (Higashinaka et al., 2015a) is mainly based on Grice's maxims of conversation (Grice, 1975), which are principles in cooperative dialogue. Grice's maxims of conversation identify the cooperative principles to be met in a general conversation between humans in terms of quantity, quality, relevance, and manner. Since the scope of a dialogue can be typically classified into utterance, response [adjacency pair (Schegloff and Sacks, 1973)], context (discourse), and environment (outside of dialogue), the taxonomy was created by combining the four maxims with the four scopes, namely, a deviation from each principle in each scope.

By eliminating invalid combinations of principle and scope (such as "relevance" and "utterance" because relevance cannot be considered for a separate utterance) and by adding system-specific errors identified through observation, 16 error types were identified for the taxonomy as shown in Table 1. The taxonomy has a main category representing the scope and a subcategory representing the deviation from Grice's maxims. For example, "Excess/lack of information" denotes the violation of the maxim of quantity in the scope of response. For further details, see (Higashinaka et al., 2015a).

The taxonomy was evaluated on the basis of inter-annotator agreement. This was done by annotating system utterances that caused dialogue breakdowns with the error types. The inter-annotator agreement was reported to be low at about 0.24 (Higashinaka et al., 2019). One of the possible reasons was the nature of human-system dialogue, which is fraught with errors, making the dialogue and the behavior of users different from those of human-human dialogue. This could have made the notions of Grice's maxims difficult to apply, leading to the low inter-annotator agreement.

Main category	Subcategory
Utterance	Syntactic error
	Semantic error
	Uninterpretable
Response	Excess/lack of information
	Non-understanding
	No relevance
	Unclear intention
	Misunderstanding
Context	Excess/lack of proposition
	Contradiction
	Non-relevant topic
	Unclear relation
	Topic switch error
Environment	Lack of common ground
	Lack of common sense
	Lack of sociality

Table 1: Theory-driven taxonomy

2.2 Data-driven taxonomy

The data-driven taxonomy (Higashinaka et al., 2015b) was created by clustering comments (textual descriptions) that describe errors made by chat-oriented dialogue systems. The comments were written by researchers working on dialogue systems. Since the number of clusters is difficult to know in advance, a non-parametric Bayesian method called the "Chinese restaurant process" (CRP) was used as a clustering method; CRP can infer the number of clusters automatically from data (Pitman, 1995). By clustering over 1,500 comments, 17 clusters were found, leading to the same number of error types. Table 2 shows the data-driven taxonomy. The names of the error types were made on the basis of observing the comments in each cluster.

The taxonomy was evaluated on the basis of the inter-annotator agreement (Higashinaka et al., 2019), in which it was found that the kappa was better than that of the theory-driven taxonomy, by which the authors concluded that it was better to use the data-driven taxonomy instead of the theory-driven one. However, there is a significant problem with the data-driven taxonomy, which is that it is too dependent on the data under analysis. The categories obtained are those brought about by the analysis of dialogue systems at a particular technical stage. The taxonomy may not be able to cope with new types of errors that may arise as a result of future development.

Category
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
Social error

Table 2: Data-driven taxonomy

2.3 Integration of taxonomies

On the basis of our observations in the previous section, we decided to integrate the two taxonomies in order to create a comprehensive one because each has shortcomings that can be covered by the other; the theory-driven taxonomy is weak in handling human-system dialogue, but the data-driven taxonomy can appropriately handle such dialogue. In contrast, the theory-driven taxonomy may cover more comprehensive dialogue phenomena on the basis of dialogue theories.

First, we decided to expand the theory-driven taxonomy to facilitate the annotation of human-system dialogue. Since system errors often deviate from the form of dialogue entirely, making Grice's maxims inapplicable, we added the distinction of "form" and "content," indicating whether or not utterances violate the normative form of dialogue, which frequently occurs in human-system dialogue. For the form, we use the normative form of language, adjacency pairs (Allen and Core, 1997), topic relevance, and social norms[2]. These represent the form in conversation that humans typically abide by and thus should be easy to detect and conceptualize. When an error does not exhibit a violation of form, we consider it to be a violation of content. Second, we placed the error types in the theory- and data-driven taxonomies into the frame of the theory-driven taxonomy expanded with form and content. Some error types fit the frame successfully, but some needed to be renamed, merged, or split to better fit the frame.

[2]Since we introduced social norms, we decided to change the scope of "environment" to "society" in the integrated taxonomy.

3 Integrated Taxonomy

Table 3 shows our taxonomy integrated through the process described in the previous section. We have 17 error types (I1–I17), each of which corresponds to a combination of the scope of dialogue and the violation of form or content. In what follows, we describe each error type in detail with dialogue examples mostly taken from actual human-system dialogues. The dialogues were originally in Japanese and were translated by the authors.

3.1 Utterance-level errors

3.1.1 Violation of Form

The violation of form at the utterance level indicates the violation of the form of language, i.e., the Japanese language in this work.

(I1): Uninterpretable: The utterance is not understandable. There are no recognizable words, or it is just a fragment of an utterance.

(1) Withha (Meaningless word in Japanese)

(I2): Grammatical error: The utterance is not grammatical or lacks important elements, such as necessary arguments and particles, for it to be a valid sentence.

(2) *Necchuusho ni ki wo tsuke ka
Heat stroke DAT care ACC take Q
"Do you take care against heat stroke?"

Here, "tsuke" (take) should be "tsukeru" or "tsukemasu" for valid Japanese conjugation.

3.1.2 Violation of Content

(I3): Semantic error: The utterance is semantically invalid such as when the combination of a predicate and its arguments cannot constitute any meaning.

(3) I am good at raining.
(one cannot be good at raining)

(I4): Wrong information: The utterance contains information that is clearly wrong to the extent that it is difficult to interpret its meaning.

(4) Bob Dylan, who is a member of the Beatles, has released a new album.
(Bob Dylan is not a member of the Beatles.)

	Violation of Form	Violation of Content
Utterance	(I1) Uninterpretable (I2) Grammatical error	(I3) Semantic error (I4) Wrong information
Response	(I5) Ignore question (I6) Ignore request (I7) Ignore proposal (I8) Ignore greeting	(I9) Ignore expectation
Context	(I10) Unclear intention (I11) Topic transition error (I12) Lack of information	(I13) Self-contradiction (I14) Contradiction (I15) Repetition
Society	(I16) Lack of sociality	(I17) Lack of common sense

Table 3: Integrated taxonomy of errors

3.2 Response-level errors

3.2.1 Violation of Form

The violation of form at the response level indicates that the utterance does not exhibit an appropriate backward-looking function for a user's previous utterance when it has a forward-looking function represented mainly by a question, request, proposal, and greeting.

(I5): Ignore question: The utterance ignores a user's question. More specifically, the utterance does not contain the content that matches the answer type requested by the question. The answer types include yes/no, agree/disagree, selection from a list, types denoted by the WH interrogatives, and entity types (person, place, etc.) requested by the question.

(5) U: What do you eat?
S: I like sports.

(6) U: In which prefecture is Yokosuka?
S: I'm moving to Yokosuka from Yokohama.

(I6): Ignore request: The utterance ignores a user's request to do something.

(7) U: Please buy it next time.
S: The costume is made to fit the hot summer, isn't it?

(I7): Ignore proposal: The utterance ignores a user's proposal/offer to do something.

(8) U: Let's talk about hobbies.
S: Which do you think is better, Urabandai (tourist location in Japan) or Taiwan?

(I8): Ignore greeting: The utterance ignores a user's greeting.

(9) U: I will go then.
S: Hello. (The system utters a greeting for opening instead of closing.)

3.2.2 Violation of Content

The violation of content at the response level indicates that, even though the surface form of the adjacency pair is appropriate, the utterance may cause a dialogue breakdown.

(I9): Ignore expectation: The utterance contains an appropriate backward-looking function for a user's previous forward-looking function; however, the utterance does not have the expected content if the underlying intention has been successfully conveyed.

(10) U: Do you have favorite sweets?
S: Yes. (The system should tell the user the name of its favorite sweets.)

Although the above example may seem to be a perfectly valid utterance, we observe cases where such a response can be inappropriate in some cooperative contexts.

3.3 Context-level errors

3.3.1 Violation of Form

We define the form for the context level as topic relevance, that is, the continuation of topics. When the system utterance is off-topic or has unclear/missing elements that keep it from being relevant to the current topics, the form is considered to be violated.

(I10): Unclear intention: Although the utterance is on a relevant topic, it does not exhibit underlying intentions (i.e., why it is mentioned) for it to be relevant. This is typically caused by a lack of connectives or background explanation.

(11) U: Hello, I like strawberries. And you?
S: I like apples.
U: I like the color red.
S: Strawberries are delicious.
(It is not clear why the system suddenly mentions strawberries although it said it liked apples.)

(I11): Topic transition error: The utterance transitions to another topic without reasonable explanation. This error type includes bringing up previous topics without reason. General top-level categories, such as food, fashion, and travel, are used as the granularity of topics.

(12) S: Oh, I love the clarinet.
U: It was really hard, but I liked it.
S: Well, what bread do you like?

When adjacent utterances do not have clear topics, this error type does not apply. In addition, when a topic transition occurs after an utterance with a backward-looking function that acknowledges the previous topic, this error type does not apply. For example, the following is not an example of a topic transition error.

(13) S: Do you plan to watch the Olympics on TV?
U: Yes, I want to.
S: On a day like today, one wants to eat hot pot for dinner.

(I12): Lack of information: The utterance misses important pieces of information, such as the subject, object, and modifier, for it to be relevant to current topics.

(14) U: Good morning! It's been nice and cool recently.
S: I like. (missing object for a predicate)

(15) U: It's too expensive, isn't it?
S: The difference is terrible. ("difference" needs things being compared)

3.3.2 Violation of Content

The violation of content at the context level indicates inconsistency or redundancy among utterances.

(I13): Self-contradiction: The utterance contradicts what has been said by that speaker.

(16) U: Where are you from?
S: I'm from Chita in Aichi. It is on the Chita peninsula.
U: Oh, Aichi. It's a nice place.
S: I just joined a company in Nagoya (city in Aichi prefecture).
U: What kind of job?
S: I'm a house wife.
(contradicts with joining a company in Nagoya)

(I14): Contradiction: The utterance contradicts or denies what has been said by the conversational partner.

(17) S: Which do you like, central league or pacific league?
U: I like central league.
S: Oh, you don't like central league.

(I15): Repetition: The utterance repeats the previous utterance (including paraphrases). This error type applies to repeated questioning.

(18) S: I'm looking forward to watching a movie.
U: Oh, yes?
S: Yeah.
U: What else do you like?
S: I look forward to watching a movie.

3.4 Society-level errors

3.4.1 Violation of Form

The violation of form at the society level indicates that the utterance is violating a social norm in a conversation, for example, being offensive to the user.

(I16): Lack of sociality: The utterance lacks consideration toward the conversational partner or particular groups of people (i.e., hate speech).

(19) U: The other day, I went to XX (country name).
S: People in XX are foolish, aren't they?

3.4.2 Violation of Content

The violation of content at the society level indicates that although the utterance is not intended to offend the user, its content is generally unacceptable.

(I17): Lack of common sense: The utterance lacks common sense. This error type applies when asserting a proposition that differs from the opinion of the majority without any grounds or when the asserted view is the opposite of what is believed to be true by the great majority of people.

(20) U: Do you want to talk about heat stroke?
S: Heat stroke is good, isn't it?

4 Evaluation

We evaluated the integrated taxonomy by annotating dialogues with error types and calculating the inter-annotator agreement. The same dialogues were annotated with the theory- and data-driven taxonomies by the same annotators for comparison.

4.1 Procedure

We used the datasets collected in past dialogue breakdown detection challenges (DBDCs), i.e., DBDC and DBDC2 (Higashinaka et al., 2016, 2017)[3], for annotating error types to system utterances that caused dialogue breakdowns. In the datasets, each system utterance was labeled with dialogue breakdown labels (B: breakdown, PB: possible breakdown, and NB: not a breakdown) by 30 annotators. We picked system utterances that were deemed inappropriate by more than a half of the annotators, that is, annotated with 15 or more B or PB dialogue breakdown labels. The dialogues were those conducted between each of three chat-oriented dialogue systems [DCM (Onishi and Yoshimura, 2014), DIT (Tsukahara and Uchiumi, 2015), and IRS (Ritter et al., 2011)] and human users. Having dialogues from multiple dialogue systems allow us to verify the applicability and coverage of our taxonomy. All dialogues were in Japanese.

There were 400 dialogues in total across the datasets. We divided the datasets into five subsets, A–E, each containing 80 dialogues. We used subsets A–C to come up with how to integrate the taxonomies. We used subset D for evaluation. We did not use subset E, which was spared for future evaluation. In the 80 dialogues, there were 599 system utterances used as a target for our error-type annotation.

[3] https://dbd-challenge.github.io/dbdc3/datasets

We annotated the error types by employing two groups of annotators. One consisted of two experts in language-annotation tasks, and the other consisted of ten crowd workers, six females and four males in their 20's to 50's. They were all certified workers of a crowdsourcing service[4] in Japan. All annotators were native Japanese. The rationale for employing the crowd workers was to ensure that the concepts of the error types were well conceptualized and easy for non-experts to understand.

All annotators performed multi-label annotation with the proposed taxonomy as well as the theory- and data-driven taxonomies. Here, since some of the error types in the data-driven taxonomy were regarded as difficult to annotate due to the ambiguity or reliance on one's understanding of dialogue systems as suggested in (Higashinaka et al., 2019), we removed "General quality," "Analysis failure," and "Mismatch in conversation" from the error types of the data-driven taxonomy. We also merged "Expression error" and "Word usage error," which were conceptually close. As a result, we had 16 and 13 error types for the theory- and data-driven taxonomies, respectively. The annotators read annotation manuals containing definitions of the error types with examples and annotated the error types on spreadsheets.

4.2 Metric for inter-annotator agreement

We used Fleiss' κ coefficient (Fleiss and Cohen, 1973) as a measure for inter-annotator agreement. Following (Ravenscroft et al., 2016), who calculated the weighted Cohen's kappa, we devised a way to calculate the weighted Fleiss' kappa. The weighted inter-annotator agreement rate P_a, extended for multi-label annotation, is calculated by,

$$P_a = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{c=1}^{C}\sum_{(l,l')} w_{ncl}w_{ncl'}}{\sum_{c=1}^{C}\sum_{(l,l')}({w_{cnl}}^2 + {w_{cnl'}}^2)/2}, \quad (1)$$

where w_{ncl} is the weight of error type c for target utterance n labeled by annotator l, N is the total number of targets for annotation, C is the number of error types, and the summation $\sum_{(l,l')}$ is taken over all combinations of annotator pairs. Note that the weights are non-negative and normalized as $\sum_{c=1}^{C} w_{ncl} = 1$. In this paper, we assume that the weights are equally distributed among the error types assigned to a target utterance. The weighted Fleiss' κ coefficient is calcu-

[4] https://www.lancers.jp/

	Experts	Crowd workers
Theory-driven taxonomy	0.186	0.206
Data-driven taxonomy	0.362	0.427
Integrated taxonomy (Proposed)	**0.567**	**0.488**

Table 4: Weighted Fleiss's κ coefficient for theory-driven, data-driven, and integrated taxonomy (proposed) by expert annotators and crowd workers.

lated by $\kappa = (P_a - P_\epsilon)/(1 - P_\epsilon)$, where

$$P_\epsilon = \sum_{c=1}^{C} \left(\frac{1}{NL} \sum_{n=1}^{N} \sum_{l=1}^{L} w_{ncl} \right)^2, \qquad (2)$$

and L is the number of annotators. The weighted agreement and Fleiss' κ coefficient are reduced to the standard ones when one of the weights is 1.

4.3 Results

The weighted Fleiss' kappa for the annotations is shown in Table 4. We can see that the agreement was higher for the integrated taxonomy compared with the theory- and data-driven ones, with reasonable kappa values of 0.576 and 0.488 for the experts and crowd workers, respectively. This result indicates that our integrated taxonomy is effective.

Using the annotations by the crowd workers, we counted the number of target utterances for which five (a half) or more annotators agreed or disagreed on the set of error types. When using the proposed taxonomy, we found that, out of 599 utterances, there were 507 utterances on which they agreed and 92 utterances on which they disagreed.

When using the theory-driven taxonomy, for the same 599 utterances, there were 126 utterances on which the annotators agreed and 473 utterances on which they disagreed. By using the proposed taxonomy, within the 473 utterances, 396 of them turned into those on which the annotators could agree. Our analysis revealed that utterances that were annotated with either "Non-understanding" or "Unclear intention" came to be reliably annotated with "Ignore question." In addition, "No relevance" and "Non-relevant topic," for which guessing the dialogue scope seems difficult, came to be reliably annotated with "Unclear intention." In addition, the introduction of "Wrong information" greatly improved the inter-annotator agreement for utterances that were otherwise labeled as "No relevance," "Unclear relation," or "Lack of common sense."

When using the data-driven taxonomy, there were 347 utterances on which the annotators agreed and 252 utterances on which they disagreed. By using the proposed taxonomy, within the 252 utterances, 193 of them became those on which the annotators could agree. Similarly to the case of the theory-driven taxonomy, the introduction of "Wrong information" was successful. In addition, such error types as "Unclear intention" and "Topic-change error" came to be reliably annotated with "Unclear intention."

Figure 1 shows a confusion matrix of annotations by the crowd workers. The matrix is calculated by

$$m_{cc'} = \sum_{(l,l')} \sum_{n=1}^{N} w_{ncl} w_{nc'l'} / \binom{L}{2}, \qquad (3)$$

which is the averaged weighted count of labels where one annotator labels type c and another labels type c'. An off-diagonal element with a large value compared with its diagonal element means confusion.

From the figure, we observed some confusions between (i) (I5) Ignore question and (I10) Unclear intention, (ii) (I10) Unclear intention and (I11) Topic transition error, and (iii) (I10) Unclear intention and (I12) Lack of information.

As for the confusion pattern (i), we observed many cases where I5 and I10 were labeled together for the same utterance. It is likely that failing to answer a question can make its interpretation difficult, leading one to wonder why the system did not respond appropriately.

Confusion pattern (ii) is probably caused by individual differences in recognizing the topic in the utterances. A typical example is as follows.

(21) U: It is scary to get used to death.
S: What? Did you fall off a Tetrapod?
U: I didn't.
S: Watermelons don't grow on trees.

In this dialogue, at the last utterance, the system suddenly talks about "watermelons." Some annotators focused on the word "watermelons" and con-

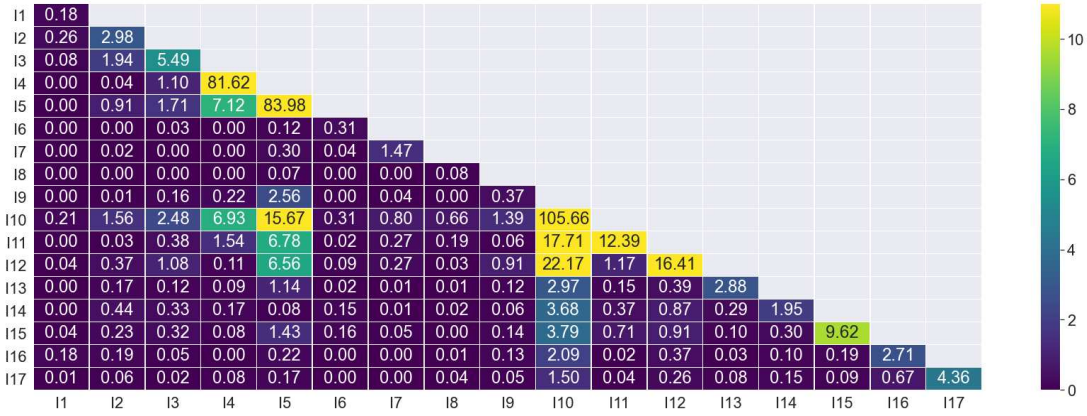

Figure 1: Confusion matrix for integrated taxonomy. I1–I17 denote error types in Table 3. Upper off-diagonal part is omitted because matrix is symmetric.

sidered the topic to be abruptly shifted. Other annotators sensed some connection in topics (probably between death and the growing of trees) but could not recognize the underlying intention.

Confusion pattern (iii) was observed for system utterances relatively short in length, which suggests that individual differences of ability in how much information can be complemented from the dialogue context cause this type of confusion. That is, if an annotator cannot complement the information, "Lack of information" is assigned; otherwise, "Unclear intention" is assigned as in the following example:

(22) U: Which do you like, domestic trips or overseas trips?
S: Travel is expensive, isn't it?
U: It depends on the tour plan. There are ones you can go on cheaply.
S: Plans sound good.

The last utterance lacks words qualifying the plans as well as explaining why or how they are "good." In such cases, it is likely that some annotators considered some expression should be added to specify the content of plans and assigned the type "Lack of information," while other annotators could not understand why "plans" are good and in what sense and used the label "Unclear intention."

4.4 Distribution of error types

Table 5 shows the distribution of error types by the proposed taxonomy for the data we used for evaluation, which includes the data of the three systems: DCM, DIT, and IRS. In addition, we also annotated the dialogues of two recently developed neural-based chatbots, Hobbyist (HBY) and ILYS-AOBA (ILA), by using two experts. For each of these two systems, we used ten dialogues that we obtained via the organizers of the dialogue system live competition that the systems were entered in (Higashinaka et al., 2020a). HBY is a Japanese version of BlenderBot (Roller et al., 2020). It utilizes 2.1B utterance pairs obtained from Twitter for pre-training and was fine-tuned by using Japanese in-house chat data (Sugiyama et al., 2020). ILA uses a similar architecture but has been trained with smaller-sized data (Fujihara et al., 2020)[5]. The two annotators first annotated dialogue breakdown labels to system utterances. Then, they performed the error-type annotation on the utterances annotated with B (breakdown) or PB (possible breakdown) labels.

	DCM	DIT	IRS	HBY	ILA
I1	0.003	0.000	0.001	0.000	0.000
I2	0.030	0.001	0.005	0.000	0.000
I3	0.044	0.013	0.005	0.000	**0.121**
I4	0.002	**0.565**	0.001	**0.300**	**0.181**
I5	**0.244**	**0.177**	**0.206**	0.014	0.036
I6	0.003	0.003	0.000	0.000	0.012
I7	0.009	0.000	0.006	0.000	0.000
I8	0.002	0.002	0.001	0.000	0.000
I9	0.012	0.002	0.018	0.067	0.061
I10	**0.334**	**0.170**	**0.458**	0.094	**0.205**
I11	0.054	0.047	**0.128**	0.028	0.072
I12	**0.130**	0.002	0.106	0.033	0.024
I13	0.023	0.004	0.011	**0.272**	0.120
I14	0.020	0.006	0.016	0.083	0.072
I15	0.052	0.008	0.016	**0.094**	0.060
I16	0.015	0.000	0.019	0.000	0.024
I17	0.025	0.001	0.003	0.014	0.012

Table 5: Distribution of error types. Three most frequent error types for each system are shown in bold.

The table shows that (I5) Ignore question and

[5] https://github.com/cl-tohoku/ILYS-aoba-chatbot

(I10) Unclear intention were frequent for DCM, DIT, and IRS, whereas there was a tendency for recent neural-based systems to suffer from (I4) Wrong information and (I13) Self-contradiction. It is interesting to see consistency in factuality and personality becoming issues in recent systems. This brief analysis shows that our taxonomy is useful for grasping error types in various chat-oriented dialogue systems.

5 Summary and Future Work

This paper proposed a new taxonomy of errors in chat-oriented dialogue systems. We integrated previously proposed theory- and data-driven taxonomies to create an integrated taxonomy. We evaluated the integrated taxonomy with Fleiss' kappa and found that our taxonomy was better than the previously proposed ones. Although there still remains some confusion between some error types, the reasonable kappa values of our taxonomy verify its validity.

As future work, we want to test the language independence because we only worked in Japanese, although we consider our taxonomy to be generally language-independent. Another possible use of the taxonomy will be to use it as a guideline for artificially generating errors so as to improve dialogue modeling in unlikelihood training (Li et al., 2019). Although the proposed taxonomy will be useful for reducing errors by systems, it will be also interesting to consider ways to recover from dialogue breakdowns after they have occurred (Higashinaka et al., 2020b). Various studies have been done on understanding how people react during miscommunication, such as by making repairs (Purver et al., 2018) and clarification requests (Liu et al., 2014; Stoyanchev et al., 2013; Rodríguez and Schlangen, 2004). We aim to expand our work to deal with various phenomena centering around dialogue breakdown. Finally, we have released the annotation manual[6] (Japanese version and its English translation) so that it can be used for the analysis of various chat-oriented dialogue systems in the community.

References

John Aberdeen and Lisa Ferro. 2003. Dialogue patterns and misunderstandings. In *Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pages 17–21.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

James Allen and Mark Core. 1997. Draft of DAMSL: dialog act markup in several layers. https://www.cs.rochester.edu/research/cisd/resources/damsl/.

Rafael E Banchs and Haizhou Li. 2012. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proc. the ACL 2012 System Demonstrations*, pages 37–42.

Niels Ole Bernsen, Hans Dybkjær, and Laila Dybkjær. 1996. Principles for the design of cooperative spoken human-machine dialogue. In *Proc. ICSLP*, volume 2, pages 729–732.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (ConvAI2). *arXiv preprint arXiv:1902.00098*.

Laila Dybkjær, Niels Ole Bernsen, and Hans Dybkjær. 1996. Grice incorporated: cooperativity in spoken dialogue. In *Proc. COLING*, volume 1, pages 328–333.

Myroslava O Dzikovska, Charles B Callaway, Elaine Farrow, Johanna D Moore, Natalie Steinhauser, and Gwendolyn Campbell. 2009. Dealing with interpretation errors in tutorial dialogue. In *Proc. SIGDIAL*, pages 38–45.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Riki Fujihara, Yosuke Kishinami, Ryuto Konno, Shiki Sato, Tasuku Sato, Shumpei Miyawaki, Takuma Kato, Jun Suzuki, and Kentaro Inui. 2020. Ilys aoba bot: A chatbot combining rules and large-scale neural response generation. *SIG-SLUD*, B5(02):110–115. (In Japanese).

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. New York: Academic Press.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2019. Improving taxonomy of errors in chat-oriented dialogue systems. In *Proc. IWSDS*, pages 331–343.

[6] https://github.com/ryuichiro-higashinaka/taxonomy-of-errors

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015a. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proc. SIGDIAL*, pages 87–95.

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. 2017. Overview of dialogue breakdown detection challenge 3. In *Proc. Dialog system technology challenge*.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proc. LREC*, pages 3146–3150.

Ryuichiro Higashinaka, Kotaro Funakoshi, Tetsuro Takahashi, Michimasa Inaba, Yuiko Tsunomori, Reina Akama, Mayumi Usami, Yoshiko Kawabata, Masahiro Mizukami, Masato Komuro, and Dolça Tellols. 2020a. The dialogue system live competition 3. *SIG-SLUD*, B5(02):96–103. (In Japanese).

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.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015b. Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proc. EMNLP*, pages 2243–2248.

Ryuichiro Higashinaka, Yuiko Tsunomori, Tetsuro Takahashi, Hiroshi Tsukahara, Masahiro Araki, Jo ao Sedoc, Rafael E. Banchs, and Luis F. D'Haro. 2020b. Overview of dialogue breakdown detection challenge 5. In *Proc. WOCHAT+DBDC5*.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2019. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. *arXiv preprint arXiv:1911.03860*.

Alex Liu, Rose Sloan, Mei-Vern Then, Svetlana Stoyanchev, Julia Hirschberg, and Elizabeth Shriberg. 2014. Detecting inappropriate clarification requests in spoken dialogue systems. In *Proc. SIGDIAL*, pages 238–242.

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

Jim Pitman. 1995. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in cognitive science*, 10(2):425–451.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: the science behind the Alexa Prize. *CoRR*, abs/1801.03604.

James Ravenscroft, Anika Oellrich, Shyamasree Saha, and Maria Liakata. 2016. Multi-label annotation in scientific articles—the multi-label cancer risk assessment corpus. In *Proc. LREC*, pages 4115–4123.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proc. EMNLP*, pages 583–593.

Kepa Joseba Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proc. SEMDIAL*, pages 101–108.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2013. Modelling human clarification strategies. In *Proc. SIGDIAL*, pages 137–141.

Hiroaki Sugiyama, Hiromi Narimatsu, Masahiro Mizukami, Tsunehiro Arimoto, Yuya Chiba, Toyomi Meguro, and Hideharu Nakajima. 2020. Development of conversational system talking about hobby using transformer-based encoder-decoder model. *SIG-SLUD*, B5(02):104–109. (In Japanese).

Hiroshi Tsukahara and Kei Uchiumi. 2015. System utterance generation by label propagation over association graph of words and utterance patterns for open-domain dialogue systems. In *Proc. PACLIC*, pages 323–331.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Richard S Wallace. 2009. The anatomy of A.L.I.C.E. In *Parsing the Turing Test*, pages 181–210. Springer.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proc. ACL*, pages 2204–2213.