Towards Taxonomy of Errors in Chat-oriented Dialogue Systems

Ryuichiro Higashinaka[1], **Kotaro Funakoshi**[2], **Masahiro Araki**[3],
Hiroshi Tsukahara[4], **Yuka Kobayashi**[5], **Masahiro Mizukami**[6]
[1]NTT Corporation, [2]Honda Research Institute Japan, [3]Kyoto Institute of Technology,
[4]Denso IT Laboratory, Inc., [5]Toshiba Corporation, [6]Nara Institute of Science and Technology

Abstract

This paper presents a taxonomy of errors in chat-oriented dialogue systems. Compared to human-human conversations and task-oriented dialogues, little is known about the errors made in chat-oriented dialogue systems. Through a data collection of chat dialogues and analyses of dialogue breakdowns, we classified errors and created a taxonomy. Although the proposed taxonomy may not be complete, this paper is the first to present a taxonomy of errors in chat-oriented dialogue systems. We also highlight the difficulty in pinpointing errors in such systems.

1 Introduction

The last decade has seen an emergence of systems that can engage in chat, small talk, or open-domain conversation. Such systems can be useful for cultivating trust between a system and users (Bickmore and Cassell, 2001), entertaining users (Wallace, 2004; Banchs and Li, 2012; Wilcock and Jokinen, 2013), and obtaining preferences from users for recommendations (Bang et al., 2015).

Error analysis is important to improve any system. However, little is known about the types of errors that can be made in chat-oriented dialogue systems. This is in contrast with many studies on task-oriented dialogue systems in which various taxonomies of errors have been proposed (Dybkjær et al., 1996; Möller et al., 2007; Ward et al., 2005; Green et al., 2006).

This paper presents a taxonomy of errors in chat-oriented dialogue systems. In our approach, we collect dialogues with a chat-oriented dialogue system and identify breakdowns (situations in which users cannot proceed with the conversation (Martinovsky and Traum, 2003)) as possible points of errors. Then, we classify the errors that led to such breakdowns to create a taxonomy. By having such a taxonomy, we hope to better grasp the main causes of breakdown in current chat-oriented dialogue systems; thereby, making it possible to make improvements. The contributions of this paper are that this is the first attempt to create a taxonomy of errors in chat-oriented dialogue systems and that we quantitatively show, by the distribution of error categories and inter-annotator agreement, the possibilities and difficulties in pinpointing errors in chat-oriented dialogue systems.

In Section 2, we cover related work on creating a taxonomy of errors in dialogue systems. In Section 3, we describe our data collection followed by the annotation of breakdowns in Section 4. In Section 5, we discuss the taxonomy we devised. In Section 6, we evaluate the taxonomy in terms of the distribution of errors and inter-annotator agreement. In Section 7, we summarize the paper and mention future work.

2 Related Work

In task-oriented dialogue systems, there is a good body of research related to the classification of errors. There are several ways to categorize errors.

One is to adopt the general taxonomy of miscommunication proposed by Clark (1996). According to Clark, there are four levels in communication; channel, signal, intention, and conversation, and by using these four levels, errors can be classified into four categories depending on which level the errors occurred. For example, if the system fails to take in audio input, it is regarded as a channel-level error. Bohus and Rudnicky (2005) applied this taxonomy to classify their non-understanding errors. A similar categorization was used by Möller et al. (2007) for their smart home and restaurant information systems. Paek (2003) discussed the generality of using the four levels for error analysis in dialogue systems, referring to the use cases across disciplines.

Proceedings of the SIGDIAL 2015 Conference, pages 87–95,
Prague, Czech Republic, 2-4 September 2015.

From the viewpoint of cooperativeness in dialogue, there is also a taxonomy based on Grice's maxims (Grice, 1975). Dybkjær et al. (1996) and Bernsen et al. (1996) had four categories of errors related to Grice's maxims; quantity, quality, relevance, and manner. They also added partner asymmetry, background knowledge, and metacommunication as error categories from their observation. Their evaluation indicated that the taxonomy can appropriately classify errors in their flight reservation system. The work by Möller (2005) also incorporated Grice's maxims into "cooperativity-related parameters" as important elements that affect interaction quality in telephone-based services.

There is also an approach to creating a task or system-specific taxonomy or errors. Aberdeen and Ferro (2003) analyzed misunderstandings by a DARPA communicator system and classified its errors into such categories as failure to obey command and repeated prompt. There is also a taxonomy for a service robot (Green et al., 2006), in which major errors are robot-specific, such as timing and reference (pointing) errors. Dzikovska et al. (2009) also classified errors in a tutorial dialogue system.

Dialogue systems are usually composed of various modules. Therefore, there is also an approach to attributing errors to modules. Ward et al. (2005) attributed causes of errors to modules, such as speech recognition, understanding, generation, and synthesis, and discussed their relative impact on usability. This approach is useful when the system has clear separation of modules.

Our approach is similar to that of (Dybkjær et al., 1996) in that we incorporate Grice's maxims into our error categories (See Section 5) and that we add other categories by our observation. The difference is that we deal with chat, which is very different from task-oriented dialogue. In this paper, we do not use their taxonomy to avoid preconception about possible errors. In this work, we did not use the four levels by Clark because we currently deal with text-based systems in which channel and signal level errors rarely occur. In addition, we do not categorize errors by modules as in (Ward et al., 2005) because chat-oriented dialogue systems usually do not have clear separation of modules.

3 Data Collection

Since there are no chat data available for analysis, we decided to create our own data set using a publicly available chat API. In this section, we describe our system based on the API, data collection procedure, and statistics of the collected dialogues.

3.1 System

We built a web-based dialogue data collection system using a chat API by NTT DOCOMO (Onishi and Yoshimura, 2014). The system is text-based, and users can chat with it in Japanese. Since the API is based on the system by Higashinaka et al. (2014a), which uses a number of natural language processing techniques (e.g., zero anaphora resolution, language generation from predicate-argument structures, and open-domain question answering), the system can be regarded as one of the most advanced (or at least complex) chat-oriented dialogue systems.

Users can chat with the system on a web browser to create 21-utterance-long dialogues; each dialogue consists of one system prompt followed by ten user-system utterance pairs. Figure 1 shows an excerpt of a dialogue collected with the system. As can be seen, the content of the conversation is open-domain. See (Higashinaka et al., 2014a) for more dialogue examples.

3.2 Procedure

We called for dialogue researchers (and their collaborators) in Japan to participate in our data collection. Since the system may not live up to their full expectations, we instructed them to follow a simple guideline that basically asked them to be as cooperative as possible. We also restricted the use of emoticons such as facial expressions. Other than that, the participants were free to chat about any topic.

3.3 Collected dialogues

One-hundred sixteen participants including the authors conversed using the system, resulting in the collection of 1,146 dialogues. Table 1 lists the statistics. This is a reasonably large data set for analyses with over 10,000 user utterances.

4 Annotation of dialogue breakdowns

Before analyzing errors, it is necessary to identify the areas of errors. To this end, we asked

S: Hello! I've been waiting for you.
U: Today, the temperature will hit 36 degrees Celsius. It's really hot.
S: Is the max temperature expected to be 17?
U: Not at all. It's a heat wave!
S: Are you prepared for a heat wave?
U: Yeah, with sunshades, water, and air-conditioners. . .
S: Air-conditioners make it comfortable.
U: Yeah, my company doesn't allow me to use air-conditioners in the morning, I'm gonna die.
S: In the company, air-conditioners really work.
U: Daytime, they work. They are necessary because we have many PCs here.
S: They don't work.
U: Really? Ours are fine here.
S: PCs are not user-friendly.

Figure 1: Excerpt of collected dialogue. S and U stand for system and user utterances, respectively. Dialogue was originally in Japanese and translated by authors.

# of Dialogues		1,146
# of Participants		116
	User	System
# of Utterances	11,460	12,606
# of Unique Utterances	10,452	7,777
# of Words	86,367	76,235
# of Unique Words	6,262	5,076

Table 1: Statistics of collected dialogues

annotators (researchers and their collaborators as in Section 3.2) to label system utterances indicating whether the utterances led to dialogue breakdowns. We used three labels depending on how easy/difficult it is to continue the conversation after each system utterance. The three labels are as follows:

(1) Not a breakdown: It is easy to continue the conversation.

(2) Possible breakdown: It is difficult to continue the conversation smoothly.

(3) Breakdown: It is difficult to continue the conversation.

We first divided the data into two sets: init100 (consisting of 100 randomly sampled dialogues) and rest1046 (the remaining 1046 dialogues). After some trial annotations with init100, we split rest1046 into eleven subsets (a–k) of 100 dialogues each (subset k contained only 46 dialogues) and allocated two annotators for each subset. For ten dialogues within each subset, we asked the annotators to provide reasons for their annotations as comments.

	Breakdown label	Ratio	Freq.
(1)	Not a breakdown	59.3%	13,363
(2)	Possible breakdown	25.3%	5,805
(3)	Breakdown	16.4%	3,752

Table 2: Distributions of breakdown annotations for rest1046 data set

Table 2 shows the distribution of the three breakdown labels for the entire rest1046 data set. This shows that we have a good percentage (about 40%) of breakdowns for analysis. The inter-annotator agreement in Fleiss' κ was 0.28 (the macro-average over the subsets), showing the subjective nature of dialogue breakdown.

5 Creating taxonomy of errors

We manually examined the system utterances annotated with breakdowns and the comments provided by the annotators to create our taxonomy of errors. After several iterations of devising error categories and annotating system utterances with the categories, we reached our agreed-upon taxonomy. We explain the taxonomy in detail as follows.

5.1 Taxonomy

Since there were many comments related to the grammar and semantics of single utterances as well as the violation of adjacency pairs (Schegloff and Sacks, 1973) and pragmatic constraints, we thought it was better to have **main categories** that distinguish to which scope of the context the errors relate; namely, we distinguished utterance-level, response-level (adjacency pair), context-level (local context (Allen, 1995)), and environment-level (not within the local context) errors.

Within each main category, we created **sub-categories**. Since there were many comments mentioning the violation of principles of cooperativeness, we created sub-categories that correspond to Grice's maxims for response and context-level errors. We then added other error categories.

We describe each error category with examples as follows.

5.1.1 Utterance-level errors

Utterance-level errors are those that can be recognized as errors by looking at single utterances. They typically come from language-generation deficiencies of the system. Note that the system generates on the basis of automatically acquired resources from the web (Higashinaka et al., 2014a). In this level, we identified three subcategories.

Syntactic error: The utterance is not grammatical or lacks important elements to be a valid sentence such as missing necessary arguments and particles.

(1) *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.

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

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

(3) The work is honest.
(work cannot be honest)

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

(4) Chu-Chu
(seemingly expressing some noise)

5.1.2 Response-level errors

Response-level errors are those that can be recognized as errors by looking at adjacency pairs. In this level, we identified five sub-categories in which the first four relate to Grice's maxims.

Excess/lack of information: The utterance violates the maxim of quantity. The utterance misses some piece of information in response to the previous user utterance or contains unnecessary information.

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

(6) U: I think simple is best.
S: Oh, it's a serious problem.
(missing reason for a proposition)

Non-understanding: The utterance violates the maxim of quality. Because of the lack of understanding, the content of the utterance is false or inappropriate in terms of the previous user utterance such as when the system says something that contradicts what the user has said.

(7) U: I'm into video games lately.
S: Do you stay away from video games?

No-relevance: The utterance violates the maxim of relation. The utterance does not have a relation to the previous user utterance such as when the system ignores a user's question and continues with other topics.

(8) U: What is your hobby, then?
S: When was your last trip?

Unclear intention: The utterance violates the maxim of manner. The utterance relates to the previous user utterance in some way but the relation is not clear, leading to the difficulty in understanding the underlying intention of the utterance.

(9) U: What catches fire?
S: Fire goes into the oven.

Misunderstanding: The utterance displays the misunderstanding of the previous user utterance such as when the system mishandles the sense of a polysemic word.

(10) U: Good evening. I think watermelons are delicious.
S: Are you shopping using a SUICA?

Watermelon ("suika" in Japanese) was misunderstood for "SUICA", an IC-card used for masstransit and shopping in Japan.

5.1.3 Context-level errors

Context-level errors are those that can be recognized as errors by looking at the local context, not just by adjacency pairs. In this level, we identified five sub-categories in which the first four relate to Grice's maxims. Compared to response-level errors, the errors here are more to do with propositions and topics in the current context.

Excess/lack of proposition: The utterance violates the maxim of quantity. The utterance does not provide any new proposition to the discourse context (e.g., repetition of already mentioned propositions) or provides excessive information than required.

(11) S: A typhoon is strong.
U: Because of it, I got soaked.
S: The force of the typhoon is strong.

Contradiction: The utterance violates the maxim of quality. The utterance contains propositions that contradict what has been said by the system or by the user. For example, the system changes its opinion or preferences within a dialogue.

(12) S: It's nice to go shopping alone.
U: I agree. That's nice.
S: Shopping takes time.
U: Window shopping is also fun.
S: It's fun to go shopping with somebody.

Non-relevant topic: The utterance violates the maxim of relation. The topic of the utterance is irrelevant to the current context such as when the system suddenly jumps to some other topic triggered by some particular word in the previous user utterance.

(13) S: I lose my SUICA.
U: Don't lose it!
S: I can get it reissued.
U: But, you can't get your deposit back.
S: I'll be careful.
U: Why don't you hang it from your neck.
S: You swing it, don't you?

'It' seems to refer to 'neck', which is a totally different topic from the previous topic "SUICA".

Unclear relation: The utterance violates the maxim of manner. Although the utterance relates to the previous user utterance, its relation to the current topic is unclear. Note that the utterance constitutes an appropriate adjacency pair but is contextually inappropriate.

(14) S: You are good at surfing.
U: I haven't tried yet.
S: The style is important.
U: Good style is cool.
S: Slender are stylish.

Although the system is talking about "style", its relation to surfing is unclear.

Topic switch error: The utterance displays the fact that the system missed the switch in topic by the user, continuing with the previous topic.

(15) S: I am going to the pool in the afternoon.
U: I envy you.
S: Oh!
U: It seems to continue to be hot in the afternoon.
S: The pool is dirty.

5.1.4 Environment-level errors

Environment-level errors are those that cannot be recognized as errors by the local context but by other issues outside the dialogue. In this level, we identified three sub-categories.

Lack of common ground: The utterance introduces, out of the blue, a proposition that has not been yet grounded.

(16) Tuna is stale, isn't it?
(Not all tuna is stale.)

Lack of common sense: The utterance displays a lack of common sense such as when the utterance has a proposition that is believed to be false in the real world.

(17) It's good to have a disease.

Lack of sociality: The utterance violates social common sense such as when the utterance breaks the social relationships with users such as insults and lack of politeness.

(18) U: I will look for a good pool for you.
S: Why don't you know it!

Main cat.	# of Ann	Ratio
UTT	220	0.126
RES	889	0.511
CON	521	0.299
ENV	110	0.063

Table 3: Number of annotations given to each main category. UTT, RES, CON, and ENV denote utterance, response, context, and environment levels, respectively.

6 Evaluation of the taxonomy

To test the validity of our taxonomy, we asked annotators to label system utterances in our data with our error categories.

One way to check the validity of a taxonomy is to observe the distribution of the annotations. When the annotations are biased towards certain categories, it is an indication that the taxonomy is not decomposing the phenomena appropriately. Another way for verifying the taxonomy is to check inter-annotator agreement. If the inter-annotator agreement is high, it is an indication that the categories are appropriately separated from each other.

We assigned three annotators for each subset of a–j (See Section 4; we did not use subset k because it had a small number of dialogues). Within each subset, we asked the annotators to annotate system utterances in the ten dialogues that had obligatory comments for breakdowns so that they could use the comments as hints to facilitate annotation. For each system utterance in question, a single error category label (i.e. sub-category label) was annotated. We instructed the annotators to check error categories from the utterance level to the environment level; that is, they first check whether the system utterance is an utterance-level error, if it is not, the check proceeds to the response level. For checking the response-level error, it was recommended that the annotators hide the context so that they can just focus on the adjacency pairs.

With this annotation process, 580 system utterances were annotated by 3 annotators with our error categories, resulting in 1740 (580 × 3) annotations. Note that we could not use the same annotators for all data because of the high burden of this annotation.

Main	Sub	# of Ann	Ratio
UTT	Syntactic error	48	0.028
	Semantic error	143	0.082
	Uninterpretable	29	0.017
RES	Excess/lack of information	185	0.106
	Non-understanding	292	0.168
	No relevance	168	0.097
	Unclear intention	186	0.107
	Misunderstanding	58	0.033
CON	Excess/lack of proposition	125	0.072
	Contradiction	132	0.076
	Non-relevant topic	71	0.041
	Unclear relation	95	0.055
	Topic switch error	98	0.056
ENV	Lack of common ground	41	0.024
	Lack of common sense	36	0.021
	Lack of sociality	33	0.019

Table 4: Number of annotations given to each subcategory. Ratio is calculated over all annotations.

6.1 Distribution of error categories

Table 3 shows the distribution of annotations summarized by the main categories. As can be seen from the table, the response-level error has the most annotations (more than 50%), followed by the context-level error. We also see quite a few utterance-level and environment-level errors.

Table 4 shows the distribution of annotations by sub-category. Within the utterance-level category, the semantic error is dominant. For the other levels, the errors seem to be equally distributed under each main category, although the number of RES-Non-understandings is larger and that of RES-Misunderstandings is less than the others. This is an indication that the taxonomy has a good categorization of errors since the distribution is not biased to only a small number of categories.

6.2 Inter-annotator agreement

Table 5 shows Fleiss' κ for main and subcategories of errors. The kappa values were calculated within each subset because the annotators were different for each subset. The average value indicates the macro-average over the subsets.

For the main categories, the averaged Fleiss' κ was 0.4, which we consider as moderate agreement. It is quite surprising that there was some difficulty in distinguishing between such obvious levels of discourse scope. For a detailed analysis, we created a confusion matrix for the main cate-

Subset	# of Utts	Main cat.	Sub cat.
a	45	0.472	0.284
b	46	0.263	0.208
c	59	0.372	0.252
d	67	0.405	0.207
e	55	0.485	0.098
f	81	0.528	0.336
g	54	0.353	0.312
h	61	0.359	0.275
i	46	0.367	0.131
j	66	0.396	0.292
Avg		0.400	0.239

Table 5: Fleiss' κ for main and sub-categories of errors. # of Utts indicates number of annotated utterances in each subset.

	UTT	RES	CON	ENV
UTT	246	140	27	27
RES	140	1242	330	66
CON	27	330	654	31
ENV	27	66	31	96

Table 6: Confusion matrix for main categories

gories (See Table 6). There was most confusion with RES vs. CON. This may be understandable because responses are closely related to the context. It is also interesting that there was much confusion regarding UTT vs. RES. Some annotators seemed to be lenient on syntactic/semantic errors and considered such errors to be response-level. Another interesting point is regarding ENV because it was most confused with RES, not CON, which is in the next level. This may be attributable to the fact that ENV is concerned with something more than the discourse scope. Although we instructed annotators to proceed from utterance-level to environment-level errors, it might have been difficult for them to ignore easy-to-find errors related to sociality and common sense.

For the sub-categories, the averaged Fleiss' κ was 0.239, which is rather low. For subset e, the kappa value was as low as 0.098. To further investigate the cause of this low agreement, we created a confusion matrix for the sub-category annotations. Since there are 16 sub-categories and the number of possible confusing pairs is 120 (${}_{16}C_2$), for brevity, we only show the top-10 confusing sub-category pairs (See Table 7). From the table, the top six pairs are all between response-level errors. The top six confusing pairs comprise about 20% of all confusions. After that, the confused pairs are mostly between response and context levels.

The confusion between RES-Non-understanding and RES-No-relevance was probably because of the perception of "what the system really understood". Some annotators thought the system made an utterance that did not match the content of the previous user utterance because it did not "understand" the user; therefore, he/she used the RES-Non-understanding category, whereas others just used the RES-No-relevance category. In fact, other confusing pairs in the response level had similar problems. For example, the category RES-Excess/lack-of-information was confused with RES-Unclear-intention because some annotators thought the intention was unclear due to the lack of information. This lack of information also made an utterance seem irrelevant in some cases.

This analysis made it clear that it is difficult to distinguish between the categories related to Grice's maxims. This may be reasonable since Grice's maxims are not claimed to be mutually exclusive. However, considering that the maxims have been successfully used to classify errors in task-oriented dialogue (Bernsen et al., 1996; Dybkjær et al., 1996), this can be due to the nature of chat-oriented dialogue systems. Our hypothesis for this confusion is that system utterances in current chat-oriented dialogue systems are far from being cooperative; thus, are not placed within the understandable regions of conversational implicature, making the classification highly subjective. Another hypothesis is that there can be multiple cooperativeness errors for the same utterance. In this case, our single-label classification scheme may not be appropriate because it necessitates the subjective choice between the cooperativeness errors.

6.3 Discussions

Since errors were not biased to particular error categories in the annotation, the taxonomy seems to have a good decomposition of errors. The main categories, which roughly distinguish the errors by the scope of discourse context, also seem to be reasonable from moderate inter-annotator agreement. However, we encountered very low inter-annotator agreement for the sub-categories. According to our analysis, it was the difficulty in distinguish-

	Confusing sub-categories		Ratio	Accum
1	RES-Non-understanding	RES-No relevance	0.048	0.048
2	RES-Excess/lack of information	RES-Unclear intention	0.034	0.082
3	RES-Excess/lack of information	RES-Non-understanding	0.032	0.114
4	RES-Excess/lack of information	RES-No relevance	0.028	0.142
5	RES-No relevance	RES-Unclear intention	0.027	0.169
6	RES-Non-understanding	RES-Unclear intention	0.025	0.194
7	RES-Non-understanding	CON-Topic switch error	0.024	0.218
8	RES-Non-understanding	CON-Contradiction	0.017	0.235
9	CON-Non-relevant topic	CON-Unclear relation	0.017	0.252
10	RES-Unclear intention	CON-Unclear relation	0.017	0.270

Table 7: Top-10 confusing sub-category pairs

ing among the categories related to Grice's maxims that attributed to this low agreement, due to the possible reason of subjectivity.

While we improve the categories and the labeling scheme to cope with the subjectivity, our suggestion for the time being is to shrink Grice's maxim-related categories (in both RES and CON) to one "cooperativeness error" category. To support this idea, we shrank such categories and recalculated Fleiss' κ. As a result, we found that the inter-annotator agreement increased to 0.358 (macro-average over the subsets). Considering that this kappa value is bounded by that of the main categories (i.e., 0.4), the reliability of this classification is reasonable.

7 Summary and Future Work

We presented a taxonomy of errors in chat-oriented dialogue systems. Through a data collection of chat dialogues and analyses of dialogue breakdowns, we created a taxonomy of errors. We then evaluated the validity of our taxonomy from two view points: the distribution of error categories and inter-annotator agreement. We argued that our taxonomy is reasonable, although some amendments are necessary. Our contributions are that we presented the first taxonomy of errors in chat-oriented dialogue systems and quantitatively evaluated the taxonomy and highlighted the difficulties in mapping errors to Grice's maxims in such systems.

There are a number of limitations in this work. First, the kappa is still low. We need to refine the categories and their definitions to reduce subjectivity in our classification scheme. It may also be necessary to incorporate a multi-label scheme. Another limitation is that the research was conducted using a single system. Although the system we adopted had many advanced features in terms of natural language processing, for generality, we need to verify our taxonomy using data of other chat-oriented dialogue systems. Another limitation is the modality considered. We only dealt with text, whereas there are many systems based on other modalities. The research was conducted only in Japanese, which is another limitation. Although we believe our approach is language-independent, we need to verify this with systems using other languages.

Our ultimate goal is to reduce errors in chat-oriented dialogue systems. Although we strive to reduce errors ourselves, since the errors concern many aspects of conversation, we are planning to make dialogue-breakdown detection an open challenge. To this end, we have released the data[1] to the public so that researchers in the field can test their ideas for detecting breakdowns. Although there have been approaches to detecting errors in open-domain conversation, the reported accuracies are not that high (Xiang et al., 2014; Higashinaka et al., 2014b). We believe our taxonomy will be helpful for conceptualizing the errors, and the forthcoming challenge will encourage a more detailed analysis of errors in chat-oriented dialogue systems.

Acknowledgments

In Japan, there is an error analysis campaign called "Project Next NLP", and within this project, there is a dialogue task as a sub-group. This research was conducted with the collaboration of more than 32 researchers from 15 institutions in the dialogue

---

[1] The data are available at `https://sites.google.com/site/dialoguebreakdowndetection/`

task. Although the authors of this paper are those who participated in the final design of the taxonomy, we thank all members of the dialogue task for data collection, annotation, and fruitful discussions. We also thank NTT DOCOMO for letting us use their chat API for data collection.

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.

James Allen. 1995. *Natural language understanding*. Benjamin/Cummings.

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.

Jeesoo Bang, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. 2015. Example-based chat-oriented dialogue system with personalized long-term memory. In *Proc. BigComp*, pages 238–243.

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

Timothy W. Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proc. CHI*, pages 396–403.

Dan Bohus and Alexander I Rudnicky. 2005. Sorry, i didn't catch that!–an investigation of non-understanding errors and recovery strategies. In *Proc. SIGDIAL*, pages 128–143.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

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.

Anders Green, Kerstin Severinson Eklundh, Britta Wrede, and Shuyin Li. 2006. Integrating miscommunication analysis in natural language interface design for a service robot. In *Proc. IEEE/RSJ*, pages 4678–4683.

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, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014a. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pages 928–939.

Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo. 2014b. Evaluating coherence in open domain conversational systems. In *Proc. INTERSPEECH*, pages 130–133.

Bilyana Martinovsky and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pages 11–16.

Sebastian Möller, Klaus-Peter Engelbrecht, and Antti Oulasvirta. 2007. Analysis of communication failures for spoken dialogue systems. In *Proc. INTERSPEECH*, pages 134–137.

Sebastian Möller. 2005. Parameters for quantifying the interaction with spoken dialogue telephone services. In *Proc. SIGDIAL*, pages 166–177.

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

Tim Paek. 2003. Toward a taxonomy of communication errors. In *Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pages 53–58.

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

Richard S. Wallace. 2004. *The Anatomy of A.L.I.C.E.* A.L.I.C.E. Artificial Intelligence Foundation, Inc.

Nigel G Ward, Anais G Rivera, Karen Ward, and David G Novick. 2005. Root causes of lost time and user stress in a simple dialog system. In *Proc. INTERSPEECH*, pages 1565–1568.

Graham Wilcock and Kristiina Jokinen. 2013. Wikitalk human-robot interactions. In *Proc. ICMI*, pages 73–74.

Yang Xiang, Yaoyun Zhang, Xiaoqiang Zhou, Xiaolong Wang, and Yang Qin. 2014. Problematic situation analysis and automatic recognition for Chinese online conversational system. In *Proc. CLP*, pages 43–51.