Analyzing dialogue breakdowns in chat-oriented dialogue systems

Ryuichiro Higashinaka[1], *Kotaro Funakoshi*[2], *Masahiro Mizukami*[3]
Hiroshi Tsukahara[4], *Yuka Kobayashi*[5], *Masahiro Araki*[6]

[1]NTT Corporation, [2]Honda Research Institute Japan, [3]Nara Institute of Science and Technology,
[4]Denso IT Laboratory, Inc., [5]Toshiba Corporation, [6]Kyoto Institute of Technology

Abstract

Although there has been much analysis of dialogue breakdown in task-oriented dialogue systems, little attention has been paid to the same issue in chat-oriented dialogue systems, making it difficult to produce improvements. We have therefore analyzed dialogue breakdown, which we define as problematic situations in which conversational participants cannot proceed with the dialogue, in chat-oriented dialogue systems. Specifically, we collected dialogue data and annotated dialogue breakdowns with many annotators and then analyzed the system utterances that led to such breakdowns in detail. Our manual and automated analysis revealed the possible causes of dialogue breakdown in chat-oriented dialogue systems.

Index Terms: chat-oriented dialogue system, dialogue breakdown, error analysis

1. Introduction

There is currently an error analysis campaign in Japan called "Project Next NLP" that contains a dialogue task as a subgroup. This sub-group contains 32 dialogue systems researchers affiliated with 15 institutions across the country. The aim of the dialogue task is to analyze dialogue breakdown in chat-oriented dialogue systems. Here, "breakdown" refers to a situation in which users cannot proceed with the conversation [1]. This happens often in dialogue systems and is problematic because it lowers user satisfaction. Analyzing such breakdowns will enable us to develop systems that can avoid or repair such situations. The reason we focus on chat-oriented dialogue systems is that, after decades of development in task-oriented dialogue systems, chat-oriented dialogue systems are now attracting attention from both the social and entertainment spheres [2, 3, 4, 5, 6]. A number of researchers involved with the dialogue task are currently working on chat-oriented dialogue systems.

In this paper, we report our analysis of dialogue breakdown in chat dialogues. There are many ways we could go about this analysis, including use of a system's internal information or specific situations; however, we choose to analyze dialogue breakdowns on the basis of their surface forms so that the analysis can be done independently of a system's internal information and the situations in which it is involved. This way, the results will be useful for systems with different backgrounds. Note that we only deal with text chat in this paper, although in the future we want to perform similar analysis with other modalities. Our analysis here is mainly to do with the linguistic and lexical phenomena of dialogues.

In our analysis, we first collect chat dialogues. Since human-system dialogues available for research are scarce, especially in Japan, we decided to collect our own data. Then, we annotate dialogues with breakdown labels in order to identify the areas of breakdown. After that, we analyze the system utterances judged as breakdowns. We also use unsupervised clustering and sentiment analysis in order to classify breakdown-related system utterances and mine expressions that are related to dialogue breakdown.

2. Dialogue data collection

2.1. System

We built a Web-based dialogue data collection system (a Web site) using a chat API provided by NTT Docomo [7]. The system is text-based, and users can chat with it in Japanese. Since the API is based on a system by Higashinaka et al. [8], which is based on a number of natural language processing techniques, the system can be regarded as one of the most advanced chat-oriented dialogue systems in use today. Users can access the site with an ordinary Web browser and have 21-utterance-long dialogues. Each dialogue consists of one system prompt, ten user utterances, and ten system responses. The dialogue alternates with each user/system utterance. After completing the dialogue, a button appears at the bottom of the site that leads to a page where users can download the conversation log.

2.2. Procedure

We asked 116 volunteers affiliated with the dialogue task of Project Next NLP, including the authors, to have chats on the Web site and then send us the logs. Since the system might not be as sophisticated as the participants expect, we created a few guidelines, e.g., to imagine talking with a stranger (as in the Loebner prize), to be as cooperative as possible, to start each conversation anew, and to be aware that the content would be made public later (i.e., do not insult anyone or include personal information).

2.3. Collected dialogues

We obtained 1,146 dialogues in total. Figures 1 and 2 show examples of these dialogues with breakdown annotations (see Section 3.1). The statistics, shown in Table 1, indicate that the number of dialogues and utterances is comparatively large, making it possible to perform meaningful statistical analyses.

3. Analysis of breakdowns

3.1. Breakdown annotation

To recognize areas of dialogue breakdown, we annotated dialogue data with breakdown labels. For this purpose, we randomly sampled 100 dialogues (called **init100** data set) and assigned 24 annotators to annotate each system utterance in the set with one of the following three labels: not a breakdown,

	Utterance	NB	PB	B
S	Hello. I was waiting for you to talk to me.	24	0	0
U	Really?! I'm so happy to hear that.			
S	What are you doing for your health?	24	0	0
U	Well, I make it a rule to drink green juice every day.			
S	What kind of food do you like to eat?	23	1	0
U	I love pizza!! What's more, I love Korean BBQ too.			
S	Do you like eating?	24	0	0
U	Love it so much. I'd like to eat as much as I like if I would never get fat.			
S	Speaking of Korean BBQ, do you like salted ox tongues?	23	1	0
U	I like boneless short ribs better than salted ox tongues. Boneless short ribs are so good, don't you think?			
S	You eat them at BBQ restaurants, do you?	22	1	1
U	The rooms will get dirty if you cook and eat them at home, so yes, I usually go to BBQ restaurants.			
S	You really like boneless short ribs.	21	2	1
U	Yes, I love them. What kind of foods do you like?			
S	I like boneless short ribs, too.	22	1	1
U	Do you? Cow's organs are delicious as well, aren't they?			
S	Cow's organs are good for health, right?	23	1	0
U	Right. I like cow intestines.			
S	Cow's organs are tasty.	23	1	0
U	Tasty and healthy, right?			
S	Cow's organ BBQ is a really fascinating food, isn't it?	22	2	0

Figure 1: Example of a collected dialogue with few breakdowns. S and U refer to system and user utterances, respectively. Dialogue was originally in Japanese and has been translated by the authors. Columns indicate the number of breakdown annotations given to each system utterance. See Section 3.1.

possible breakdown, and breakdown. These are defined as:

(NB) Not a breakdown: It is easy to continue the conversation after the system utterance in question.

(PB) Possible breakdown: It is difficult to continue the conversation smoothly after the system utterance in question.

(B) Breakdown: It is difficult to continue the conversation at all after the system utterance in question.

Since there are 1,100 system utterances (including system prompts) in init100, we obtained 26,400 annotations of breakdown labels. Here, annotation for a dialogue was not performed by the one who performed that dialogue. This is because we wanted to collect spontaneous dialogues from participants and to have annotations produced from a third-person point of view for the analyses.

For this annotation, we used a specialized tool we developed, shown in Figure 3. System utterances can be annotated with NB, PB, or B by clicking the corresponding buttons. There is a text field beside the buttons to input arbitrary comments about the reason for annotating various utterances with PB or B labels. Utterances were annotated from the beginning of a dialogue to the end. The annotation tool can hide the utterances after the system utterance in question so that breakdown labels can be annotated only with regard to the local context. Note that, even after a breakdown occurs within a dialogue, we asked the annotators to continue labeling succeeding utterances by regarding the context that has the breakdown as given.

The annotation scheme utilized here is similar to that used by Luzzati et al. to examine the "gravity" of automatic speech recognition (ASR) errors [9]. They used low, intermediate, and high labels to annotate ASR errors. Our work is similar in that we also focus on the content of errors on the basis of subjective judgments.

	Utterance	NB	PB	B
S	Good evening. Be careful to avoid heat stroke.	24	0	0
U	Sure.			
S	O.K.	22	2	0
U	I am drinking enough water.			
S	**Do you run short of water?**	10	11	3
U	Yes, I do, if I sweat.			
S	Do you lose water if you sweat??	22	2	0
U	Yes, I do.			
S	This is so.	22	2	0
U	You'd better stay at home, as it is very hot outside.			
S	You have sweat on your face, don't you?	22	2	0
U	Indeed.			
S	So you do.	24	0	0
U	I hate to sweat.			
S	Your make-up gets smeared due to sweat, doesn't it?	24	0	0
U	I don't wear make-up.			
S	**I want make-up.**	9	10	5
U	I don't want it.			
S	**Thanks for the meal!**	1	5	18
U	Did you eat anything?			
S	**The make-up is appropriate.**	3	6	15

Figure 2: Example of a collected dialogue with several breakdowns. Utterances in bold font are possible causes of breakdowns; that is, they are annotated with more breakdown labels than with not-a-breakdown labels.

Table 1: Statistics of collected dialogues. We used MeCab for morphological analysis to obtain word boundaries.

No. of dialogues		1,146
No. of participants		116
	User	System
No. of utterances	11,460	12,606
No. of unique utterances	10,452	7,777
No. of words	86,367	76,235
No. of unique words	6,262	5,076

3.2. Distribution of breakdown labels

Table 2 shows the overall distribution of breakdown labels in init100. As seen in the table, about 60% of the utterances are regarded as being without a problem, about 20% are related to possible breakdowns, and the remaining 20% cause breakdowns.

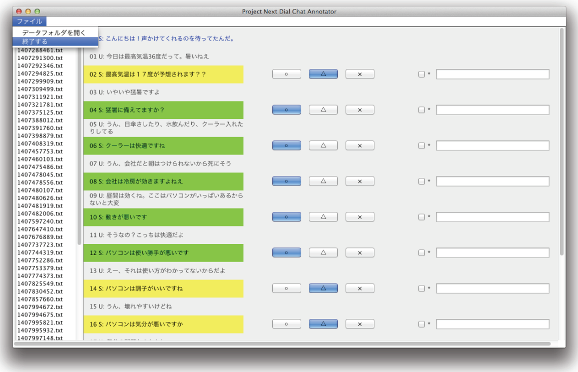

Figure 3: Annotation tool for labeling dialogue breakdowns.

Table 2: Distributions of breakdown labels in init100.

Not a breakdown	Possible breakdown	Breakdown
59.2% (14212)	22.2% (5322)	18.6% (4466)

As one would imagine, the annotation of breakdown is highly subjective. One might regard some utterances as problematic whereas others would regard them as reasonable. To address this point, we calculated Fleiss' κ between the annotators. We found that the kappa value is 0.276, which verifies the subjective nature of dialogue breakdown. When we merge PB and B to make it a two-class annotation, then, the kappa value rises to 0.396, which is moderate agreement. Among the annotators, there exists some distinction between NB and PB+B utterances. Figure 4 shows the proportion of breakdown labels per annotator, highlighting the subjective nature.

Table 3 shows the histogram of annotations classified by the number of PB, B, and PB+B. For example, the first row starting with 0 shows that the number of system utterances labeled with no PB+B labels is 205, with no PB is 217, and with no B is 400 (that is, there are 400 system utterances that were not recognized as breakdown by anyone). When we get to the bottom, we see that there are 24 system utterances with 24 PB+B labels. In other words, there are 24 system utterances that are totally problematic in init100. If the decisions of all annotators were in agreement, the histogram would only have values in the rows of 0 and 24. As we can see, this is not the case. In reality, the distribution looks rather uniform if we look at the column of PB+B.

If we look at the column of PB, there are very few utterances with 15 or more breakdown labels. In such cases, it is clear that they become B; for an utterance to be a complete breakdown point, the decision is based on whether about 60% (15/24) of the population considers it a (possible) breakdown. This is a useful insight in terms of annotation. When we have ten annotators, then, we can consider utterances annotated by more than six annotators as reliable breakdown.

3.3. Utterances with many breakdown labels

We extracted system utterances with varying numbers of breakdown labels for examination. Table 4 lists examples (Ex1 to Ex6) of system utterances that led to many breakdown labels. These are sorted by the total number of PB+B.

The first example, Ex1, has 24 labels of PB+B, which indicates that all 24 of the annotators thought this utterance led to a breakdown or possible breakdown. Here, the user's question is ignored and the system's response has no relation to the question, leading to this unanimous decision of breakdown. In Ex2, although the number of PB+B is 24, the distribution of PB and B is different from that of Ex1: it is a split case between PB and B. Here, although the system does not answer the user's question, the system's response is technically related because it is about Tokyo, which made some annotators withdraw the decision of a complete breakdown. In Ex3, the system utterance cannot be understood by the annotators at all because of the usage of slang. This indicates that when the system's utterance is not understandable, it tends to be labeled as breakdown instead of possible breakdown. In Ex4, most annotators think it is possible breakdown. In this example, the user asks about beer preference and the system replies with a drinking habit, not a direct answer to the question. The leniency of the decision comes from the fact that the answer is in some way related to the question. In Ex5, a grammatical error occurred in the system's utterance and most annotators thought this would lead to a breakdown, not a possible breakdown. In Ex6, the system suddenly reads out some news, completely ignoring the user's utterance. However, there were more PB than B breakdown labels in this case, probably because the system uses "by the way" to signal a topic switch.

A close look at the examples above provides us with some insight for avoiding breakdowns. When the system ignores a user, it is likely that breakdown will follow. However, when the system says something related to the previous utterance, the risk of breakdown is greatly reduced. When there are utterances that users cannot understand, or that contain systematic/grammatical errors, it is very likely that a breakdown will follow. In the next sections, we use statistical techniques to further mine the causes of dialogue breakdown.

Table 3: Histogram of annotations classified by the number of PB, B, and PB+B.

No. of labels	PB+B	PB	B
0	205	217	400
1	74	90	125
2	62	77	77
3	51	79	89
4	45	79	42
5	48	89	41
6	42	86	35
7	42	91	33
8	37	79	34
9	41	51	50
10	25	64	30
11	20	37	21
12	25	27	25
13	27	19	22
14	36	11	16
15	34	2	16
16	37	0	10
17	30	2	10
18	35	0	5
19	26	0	6
20	38	0	7
21	36	0	4
22	27	0	1
23	33	0	1
24	24	0	0
Total	1100	1100	1100

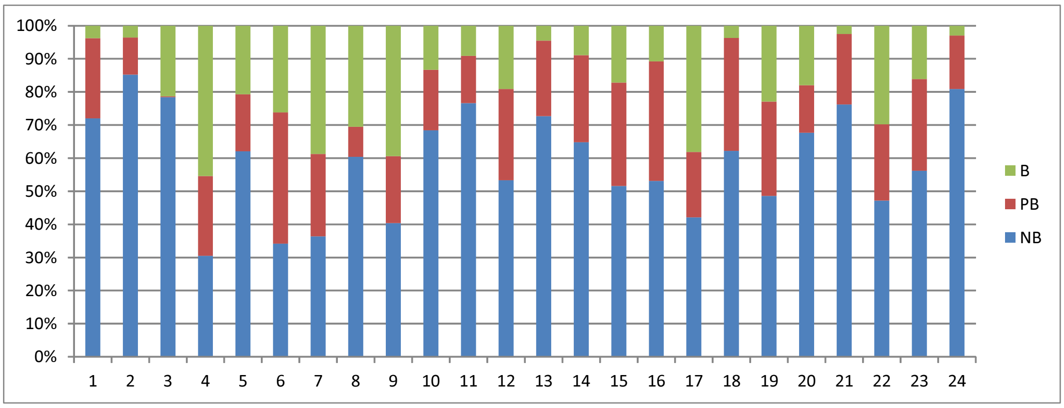

Figure 4: Proportion of breakdown labels (NB, PB, and B) per annotator. X-axis indicates annotator ID.

Table 4: Examples of system utterances with the number of breakdown labels.

	Context (previous utterance)	System utterance (that led to breakdown)	PB	B	PB+B
Ex1	U: By the way, it is winter now, isn't it?	S: I look at good-looking men.	1	23	24
Ex2	U: It is Tokyo. And you?	S: It has a large population.	13	11	24
Ex3	U: Take care.	S: Terari (a slang to say thanks like "ta" in English).	2	18	20
Ex4	U: Draft beer or bottled beer?	S: I drink beer after my bath.	15	5	20
Ex5	U: Especially in summer, it is refreshing to swim.	S: Do you learn swim? (grammatical error)	3	13	16
Ex6	U: My mom made me a meal.	S: By the way, about 70% of the reason for glacier reduction after 1991 comes from $\cdots$	13	3	16

3.4. Clustering of comments

As noted in Section 3.1, in addition to breakdown labels, annotators provided written comments about the reasons for breakdown. These comments provide rich information about the possible causes of breakdowns. One thing we can do with these comments is perform unsupervised clustering of them, deriving clusters that each encode a possible cause. Since it is difficult to know the number of clusters in advance (we do not know how many causes of errors there are in chat-oriented dialogue systems), here, we use a clustering technique called the Chinese restaurant process (CRP). CRP is a nonparametric Bayesian method that can infer the number of clusters from data. It has been successfully used in dialogue systems research for the classification of utterances [10].

Table 5 shows the results of clustering by CRP. For hyperparameters, we used 0.1 for α and β (see [10] for the details of these hyperparameters). The number for iterations was 1,000 in Gibbs sampling. We found 20 clusters and extracted representative words for each cluster. For morphological analysis, we used NTT's JTAG. For the extraction of representative words, we used log-likelihood ratio testing (similar to chi-squared testing), which uses a two-by-two matrix to test the independence of a word to a particular cluster. We list up to five words with the p-value of below 0.001. As for representative comments, we manually identified frequent comments in each cluster and summarized them in the table.

As evident in the table, each cluster seems to correspond to a particular quality feature of dialogue. The biggest cluster is about the overall quality of dialogue. This is followed by more specific points. Our interpretations of the clusters are provided in the table. The interpretations include non-understandable utterances, failure to answer questions, and ignoring users, as identified manually (see Section 3.3). The table shows many other features (errors) of dialogue that are related to dialogue breakdowns. Although we do not know as yet if this table is an exhaustive list of the causes of breakdown, it is clear that we need to consider various aspects of dialogue in chat-oriented dialogue systems from the word-level to the social-level.

Although the size of the clusters is not related to the importance of a phenomenon, since there were many comments regarding the understandability of utterances and answering questions, the generation of understandable utterances and good capability of question answering seem to be important (or at least easily noticed by users).

3.5. Sentiment analysis of comments

We performed sentiment analysis to further mine the comments. Here, we mined evaluative expressions together with their target entities (called "attributes" here). For example, if there is a comment "the meaning is not understandable", the attribute is "meaning" and its evaluative expression is "not understandable". We used our language processing tool to extract attributes and their evaluative expressions [11]. The extraction of attributes and evaluative expressions is based on hand-crafted rules. We used all the comments given to the PB and B annotations (3,329 comments in all) for this experiment.

Table 6 lists the evaluative expressions with their attributes. The attributes are sorted by frequency. It can be seen that the attributes are quite similar to those of our interpretations of the clusters in Section 3.4, which is interesting because they are the results of two different approaches. This adds credibility to the results of both CRP and sentiment analysis. Here again, the understandability of system utterances is highlighted.

The evaluative expressions for "Response" have a strong resemblance to Grice's maxims [12]; that is, "ambiguous" and "unclear" are related to the maxim of manner, "contradiction" is related to the maxim of quality, and "mismatched" is related to the maxim of relevance. Although we do not have anything for

Table 5: Clusters obtained by CRP of 1,511 comments given to breakdowns (Bs). Clusters are ordered by size (the number of comments in each cluster). Interpretation denotes our interpretation of the clusters.

ID	Size	Interpretation	Representative words	Representative comments
14	197	General quality	dialogue, well-formed, conversation	system fails to converse, conversation is not valid
8	186	Not understandable	understand, meaning, what, make sense	cannot understand meaning, cannot understand
16	152	Failure to answer questions	answer, question, partner, respond, query	does not answer question, answer is not valid, ignore questions
12	130	Contradiction	consideration, understand, utterance, before, oneself	contradicts what system said, does not consider context
9	113	Ignore user	ignore, user, utterance, contradiction, opinion	ignore user utterance, ignore user question, ignore user opinion
1	108	Unclear intention	unclear, intention, meaning, utterance	intention is not clear, meaning is not clear
6	106	Analysis failure	analysis, recognition, words, related, attention	system fails to analyze utterance
4	98	Repetition	say, add, strange, mind, tired	say things already said, say the same thing repeatedly, persistent utterance
17	82	Grammatical error	inappropriate, response, invalid, Japanese, answer	inappropriate as answer, inappropriate as Japanese, not a response
19	54	Topic-change error	change, topic, sudden, story	sudden topic change, different story, abrupt change of topic
10	52	Violation of common sense	against, response, relation	violates common sense, value does not match that of user
0	51	Expression error	different, strange, say, context	content is strange, expression is strange
15	51	Mishandling of polysemy	go, mind, Suica	fail to understand polysemic words (watermelon [suika] vs. Suica [an IC card])
3	41	Word usage error	agree, add, evaluate, use, request	invalid usage of words
13	28	(not applicable)	travel, eat	(specific errors related to the topic of travel)
5	25	Out-of-context	match, flow	does not match flow of dialogue, does not go well with context
2	24	Mismatch in conversation	mismatch, conversation	mismatch in conversation, conversation not smooth
11	6	Mismatch in response	mismatch, response	mismatch in response, something missing in response
7	5	Social error	(no words)	socially invalid, ungrammatical
18	2	No information	(no words)	no new information

the maxim of quantity, we do have "persistent" for "Repetition". We leave it as our future work to perform analyses from the viewpoint of Grice's maxims.

4. Summary and future work

In order to improve chat-oriented dialogue systems, in this paper we analyzed breakdowns in chat-oriented dialogues. We collected chat dialogues and annotated dialogue breakdowns using three labels: not a breakdown, possible breakdown, and breakdown. The system utterances were then manually analyzed, and unsupervised clustering and sentiment analyses were applied to the comments given to breakdowns in order to identify causes of breakdown. Our findings are summarized as follows:

- The reasoning behind what caused a breakdown is highly subjective, but when more than 60% of annotators agree on a breakdown, we can assume it to be a well-grounded breakdown.
- Utterances that cannot be understood by users are highly likely to lead to breakdowns.
- Relevance to context is important in differentiating between breakdowns and possible breakdowns.
- From the number of clusters and the number of attributes we found in the sentiment analysis, the possible number of causes for dialogue breakdown is around 15 to 20.
- Dialogue breakdown has a strong relation to Grice's maxims.

As future work, we would like to create a taxonomy of dialogue breakdowns so that researchers in the field can better handle and avoid dialogue breakdown. Many taxonomies of errors/miscommunication in dialogue have been created, especially in task-oriented dialogue systems [13, 14, 15]. Some of these taxonomies are related to Grice's maxims [16]. We want to use these earlier insights in creating our own taxonomy. We also want to perform analyses using other systems and systems with other languages for generality.

Our ultimate goal is to create chat-oriented dialogue systems in which as few breakdowns occur as possible. To this end, we are planning to open our dataset to the public and hold an evaluation workshop (such as the Text REtrieval Conference [17] and Dialogue State Tracking Challenge [18]) for detecting dialogue breakdowns. Although there have been approaches to detecting errors in open-domain conversation, the reported accuracies are low [19, 20], and we feel a strong need for a challenge. As evident from the findings presented in this paper,

Table 6: Evaluative expressions with attributes. Attributes with more than five occurrences are listed.

Attribute	Count	Evaluative expressions with frequencies
Meaning	129	cannot understand: 82, do not understand well: 18, unable to understand: 9, unclear: 5, cannot take in: 3, negative: 2, could not understand: 2
Response	74	ambiguous: 31, unclear: 14, mismatched: 7, contradiction: 3, a little mismatched: 3
Utterance	42	unclear: 25, contradiction: 6, contradicted: 3, not necessarily true: 2, ambiguous: 2
Intention	41	unclear: 39, strange: 2
Answer	40	mismatched: 25, a little mismatched: 3, unclear: 3, ambiguous: 2
Question	21	strange: 11
Expression	19	bad: 7, impolite: 4, a little strange: 3, circumlocutory: 3
Partner	16	strange: 6, mind: 2, sorry: 2
Intention	16	cannot understand: 11, unable to understand: 2
Content	14	mismatched: 4, contradiction: 2, strange: 2
Flow	14	unclear: 4
Topic	11	abrupt: 2
Usage	9	strange: 6, unnatural: 2
Repetition	7	persistent: 2

it is not easy to make dialogue breakdown detection a shared task because of the subjectivity. A careful consideration of the evaluation method will be crucial. As the systems improve, errors will become different. We want to cultivate a robust cycle of system development and error analysis through evaluation workshops.

5. Acknowledgments

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.

6. References

[1] B. Martinovsky and D. Traum, "The error is the clue: Breakdown in human-machine interaction," in *Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems*, 2003, pp. 11–16.

[2] T. W. Bickmore and J. Cassell, "Relational agents: a model and implementation of building user trust," in *Proc. CHI*, 2001, pp. 396–403.

[3] R. S. Wallace, *The Anatomy of A.L.I.C.E.* A.L.I.C.E. Artificial Intelligence Foundation, Inc., 2004.

[4] R. E. Banchs and H. Li, "IRIS: a chat-oriented dialogue system based on the vector space model," in *Proc. the ACL 2012 System Demonstrations*, 2012, pp. 37–42.

[5] G. Wilcock and K. Jokinen, "Wikitalk human-robot interactions," in *Proc. ICMI*, 2013, pp. 73–74.

[6] J. Bang, H. Noh, Y. Kim, and G. G. Lee, "Example-based chat-oriented dialogue system with personalized long-term memory," in *Proc. BigComp*, 2015, pp. 238–243.

[7] K. Onishi and T. Yoshimura, "Casual conversation technology achieving natural dialog with computers," *NTT DOCOMO Technical Jouranl*, vol. 15, no. 4, pp. 16–21, 2014.

[8] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proc. COLING*, 2014, pp. 928–939.

[9] D. Luzzati, C. Grouin, I. Vasilescu, M. Adda-Decker, E. Bilinski, N. Camelin, J. Kahn, C. Lailler, L. Lamel, and S. Rosset, "Human annotation of ASR error regions: Is "gravity" a sharable concept for human annotators?" in *Proc. LREC*, 2014, pp. 3050–3056.

[10] R. Higashinaka, N. Kawamae, K. Sadamitsu, Y. Minami, T. Meguro, K. Dohsaka, and H. Inagaki, "Unsupervised clustering of utterances using non-parametric bayesian methods." in *Proc. INTERSPEECH*, 2011, pp. 2081–2084.

[11] H. Asano, T. Hirano, N. Kobayashi, and Y. Matsuo, "Subjective information indexing technology analyzing word-of-mouth content on the web," *NTT Technical Review*, vol. 6, no. 9, pp. 6–13, 2008.

[12] H. P. Grice, "Logic and conversation," in *Syntax and Semantics 3: Speech Acts*, P. Cole and J. Morgan, Eds. New York: Academic Press, 1975, pp. 41–58.

[13] T. Paek, "Toward a taxonomy of communication errors," in *Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems*, 2003, pp. 53–58.

[14] D. Bohus and A. I. Rudnicky, "Sorry, i didn't catch that!– an investigation of non-understanding errors and recovery strategies," in *Proc. SIGDIAL*, 2005, pp. 128–143.

[15] S. Möller, K.-P. Engelbrecht, and A. Oulasvirta, "Analysis of communication failures for spoken dialogue systems," in *Proc. INTERSPEECH*, 2007, pp. 134–137.

[16] N. O. Bernsen, H. Dybkjaer, and L. Dybkjaer, "Principles for the design of cooperative spoken human-machine dialogue," in *Proc. ICSLP*, vol. 2, 1996, pp. 729–732.

[17] E. M. Voorhees and D. Harman, "Overview of TREC 2001," in *Proc. TREC*, 2001.

[18] J. Williams, A. Raux, D. Ramachandran, and A. Black, "The dialog state tracking challenge," in *Proc. SIGDIAL*, 2013, pp. 404–413.

[19] Y. Xiang, Y. Zhang, X. Zhou, X. Wang, and Y. Qin, "Problematic situation analysis and automatic recognition for Chinese online conversational system," in *Proc. CLP*, 2014, pp. 43–51.

[20] R. Higashinaka, T. Meguro, K. Imamura, H. Sugiyama, T. Makino, and Y. Matsuo, "Evaluating coherence in open domain conversational systems," in *Proc. INTERSPEECH*, 2014, pp. 130–133.