Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems

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

Abstract

This paper aims to find errors that lead to dialogue breakdowns in chat-oriented dialogue systems. We collected chat dialogue data, annotated them with dialogue breakdown labels, and collected comments describing the error that led to the breakdown. By mining the comments, we first identified error types. Then, we calculated the correlation between an error type and the degree of dialogue breakdown it incurred, quantifying its impact on dialogue breakdown. This is the first study to quantitatively analyze error types and their effect in chat-oriented dialogue systems.

1 Introduction

Chat-oriented or open-domain dialogue systems have recently been attracting attention from social and entertainment aspects (Bickmore and Cassell, 2001; Banchs and Li, 2012; Wilcock and Jokinen, 2013). However, since they need to deal with open-domain utterances, which current natural language processing techniques are not mature enough to handle appropriately, the system inevitably makes errors. This discourages users from talking to the system, leading to dialogue breakdowns in conversation (Martinovsky and Traum, 2003). Here, dialogue breakdowns denote points in dialogue where users are unable to continue the conversation.

This paper aims to find errors that lead to dialogue breakdowns in chat-oriented dialogue systems. Our approach is two-fold: (1) identify error types in chat-oriented dialogue systems, and (2) calculate their impact on dialogue breakdown. For (1), we first collect chat dialogues between an automated system and users, annotate the dialogues with dialogue breakdown labels, and collect comments that describe the error that led to the breakdowns. After that, we apply automatic clustering methods to the comments to obtain clusters of comments, each of which corresponds to a particular error type. For (2), we calculate the correlation between an error type and the degree of dialogue breakdown it incurs. This way, we can quantify the impact of an error type on dialogue breakdown. By our approach, we hope to distinguish fatal errors from non-fatal ones, providing useful guidelines for dialogue system developers to avoid breakdowns in chat-oriented dialogue systems.

2 Data Collection

For data collection, we asked dialogue researchers and their collaborators in Japan to use our chat-oriented dialogue system. The system is text-based and enables chatting in Japanese. It was built by wrapping a chat API provided by NTT DOCOMO (Onishi and Yoshimura, 2014). Since the API works on the basis of (Higashinaka et al., 2014), which uses a number of natural language processing techniques to understand and generate utterances, we expected to obtain a wide variety of dialogues, and hence, a variety of errors.

A total of 116 users chatted with the system, resulting in 1,146 dialogues. Here, each dialogue was controlled to be 21 utterances long: one system prompt with ten utterances each from the user and system. Then, we randomly sampled 100 dialogues (called the *init100* data set) for dialogue breakdown annotation. Twenty-four annotators subjectively labeled each system utterance in init100 with one of the following three dialogue breakdown labels:

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

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

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2243–2248,
Lisbon, Portugal, 17-21 September 2015.

	Utterance	NB	PB	B	C	Comments
S0	Hello! Talk to me anytime.	24	0	0	0	
U1	It's hot today, isn't it?					
S1	Please tell me your favorite movie genre.	20	2	2	2	User utterance is not considered. / The system ignored user's intended topic.
U2	I like action movies.					
S2	Action movies, I like them.	21	3	0	0	
U3	What kind of movies do you like?					
S3	I like Japanese ones.	23	1	0	0	
U4	That's great.					
S4	Thank you.	24	0	0	0	
U5	Do you know what movie will be aired on Friday night?					
S5	Yes, yes.	4	11	9	4	The system did not mention a movie name. / It answered only "yes" but without saying a movie name. / This is not an answer to the question. / The conversation is not valid.
U6	I haven't heard of that movie.					
S6	That's good.	3	10	11	4	This is not a response to a question. / Strange answer. / It is not clear what is good. / The conversation is not valid.
U7	Oh, is it good? I will watch it then.					
S7	That's good!	18	3	3	1	It is not clear what is good.
U8	Recently, I have had lots of free time.					
S8	That's great.	15	7	2	0	
U9	Are you busy?					
S9	Indeed so.	24	0	0	0	
U10	Take care. Have some rest.					
S10	I see.	10	9	5	2	The system should thank the user. / Strange answer.

Figure 1: Example dialogue with the number of breakdown labels for each system utterance. S and U stand for system and user utterances, respectively. C stands for the number of comments provided. The dialogue was originally in Japanese and was translated by the authors.

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

The annotators were instructed to provide voluntary comments to describe the errors that led to the breakdowns. Table 1 shows the statistics of the dialogue breakdown labels and comments. Figure 1 shows an example dialogue with the number of breakdown labels and comments for each system utterance. In this example, S5 and S6 were annotated with nine and eleven breakdown labels, respectively, both having four comments.

The inter-annotator agreement of dialogue breakdown annotation in Fleiss' kappa was 0.276, which seems relatively low. One reason for this is obviously the subjective nature of the task. Another possible reason is that we intentionally did not set rigid guidelines for dialogue breakdown annotation so as to explore possible error types in chat-oriented dialogue systems. When we merge PB and B and make it a two-class annotation, the agreement becomes 0.396 (moderate agreement), showing that the subjects share some common conception about dialogue breakdown.

Breakdown label	# of labels	# of comments
NB	14,212	57
PB	5,322	1,818
B	4,466	1,511

Table 1: Statistics related to breakdown labels and comments in init100 data set. Note that NB also had some opinions as comments.

3 Analysis

3.1 Automatic clustering of comments

We first need to identify the error types in chat-oriented dialogue systems. For this, we applied an automatic clustering method to the comments to obtain clusters of comments. Our idea is that, since each comment describes a particular error that led to a breakdown, a cluster of comments is likely to represent an error type. Since the number of clusters is difficult to know in advance, we turn to a non-parametric Bayesian method called the Chinese restaurant process (CRP) as a clustering method. CRP can infer the number of clusters automatically from data (Pitman, 1995).

We applied CRP to the 1,511 comments given

ID	Size	Interpretation	Representative words in the cluster
2	259	General quality	dialogue, well-formed, consideration, conversation
0	194	Not understandable	understand, meaning, what
6	148	Ignore user utterance	ignore, user, utterance
7	134	Ignore user question	answer, question, partner, respond
1	113	Unclear intention	unclear, intention, meaning, utterance
8	107	Contradiction	doubtful, change, negative (as opposed to positive), previous
16	100	Analysis failure	analysis, recognition, related, understand
4	95	Inappropriate answer	response, inappropriate, invalid, answer
5	77	Repetition	say, add, mind, strange, tired
3	54	Grammatical error	(words specific to particular grammar usage)
14	53	Expression error	(words specific to particular expressions)
11	39	Topic-change error	change, topic, sudden, mismatch, response
12	38	Violation of common sense	match, flow, common sense, against, connection
13	36	Word usage error	(words specific to particular word usage)
10	35	Diversion	story, different
15	25	Mismatch in conversation	match, story
9	4	Social error	(no words)

Table 2: Clusters by CRP of 1,511 comments given to breakdowns. The clusters are ordered by size.

to breakdown labels (B labels). For the clustering, we used the same procedure as (Higashinaka et al., 2011); each datum (comment) was represented by a word frequency vector, and the probability that it belonged to a particular cluster was calculated by using the likelihood that the words are generated from the word distribution of that cluster. The hyper-parameters α and β were both set to 0.1 and the number of iterations for Gibbs sampling was 10,000. See (Higashinaka et al., 2011) for the details of the procedure.

Table 2 shows the clustering results. We obtained 17 clusters. For each cluster, we mined representative words by a log-likelihood ratio test, which uses a two-by-two matrix to test the independence of a word to a particular cluster. By looking at the representative words and also the raw comments, we came up with the interpretations of the clusters as indicated in the table. Although we do not go into the details of the clusters one by one, each cluster seems to successfully represent a certain error type in chat-oriented dialogue systems. We also applied CRP to the 3,329 comments given to PB and B to obtain similar clusters except that we additionally had clusters whose interpretations are as follows: inability to handle invalid user input, missing topic, missing information, mismatch in response, no reaction, and no information. They account for about 13.3% of the comments and mostly concern missing elements (such as missing arguments) in dialogue. Since such missing elements can be complemented by follow-on utterances in dialogue, they only appear in the comments for PB; they do not lead to an immediate dialogue breakdown.

To further categorize the clusters, we applied a hierarchical clustering (an agglomerative clustering) to the clusters. Here, a cluster was represented by the word frequency vector of all comments contained in the cluster, and the similarity of the clusters was calculated by cosine similarity of word frequency vectors. For the linkage criterion, we used Ward's method. Figure 2 shows the clustering results. The figure indicates that there are the following eight main error categories (E1–E8):

(E1) Clusters 2 and 16 concern the general ability of a system.

(E2) Clusters 7, 5, and 8 relate to context awareness: the ability to recognize when it is asked a question and to recognize what has been said before.

(E3) Clusters 13, 3, and 14 concern the language generation (surface realization) ability.

(E4) Clusters 4 and 6 concern the response ability: the ability to answer questions and to create utterances relevant to the previous user utterance.

(E5) Cluster 1 relates to the exhibition of an intention or a plan: the ability to make clear the

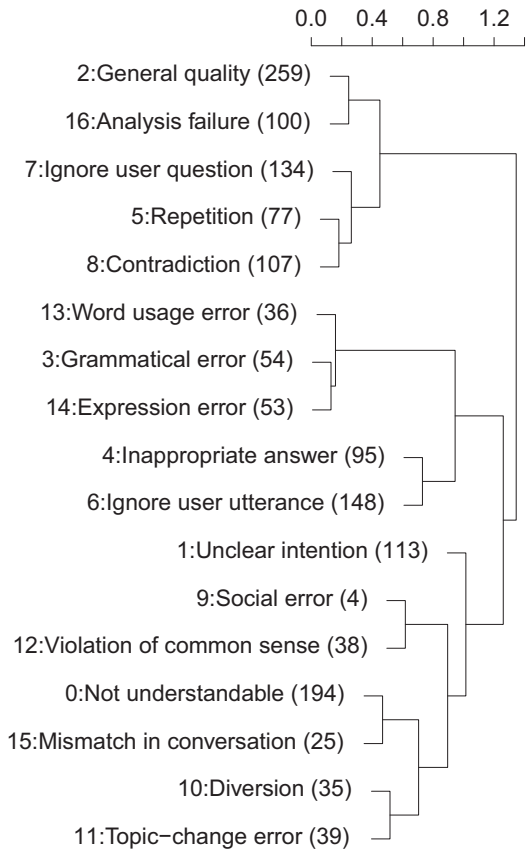

Figure 2: Hierarchical clustering applied to the obtained clusters. The numbers in parentheses denote cluster size.

purpose of an utterance.

(E6) Clusters 9 and 12 relate to the social ability: the ability not to offend users or say things that are not socially acceptable.

(E7) Clusters 0 and 15 concern the understandability of an utterance: the ability to generate utterances that have clear meanings in the context of the conversation.

(E8) Clusters 10 and 11 relate to the awareness of current topics.

3.2 Analyzing the impact of error types

Having identified the error types and error categories, we investigated their impact on dialogue breakdown. For this purpose, we examined the correlation between an error type and its degree of breakdown: the higher the correlation, the more it is related to dialogue breakdown. Specifically, we calculated the correlation ratio (η) between the existence of a comment belonging to a particular cluster (error type) and the number of breakdown labels (B labels). Note that the correlation ratio is equivalent to Pearson's correlation coefficient except that it can be applied to categorical data. The η ranges from 0 to 1. For calculation, we first extracted data that had one or more B labels and one or more corresponding comments (we had 556 such samples in our data). Then, we calculated the correlation ratios.

ID	Interpretation	Cat	η
0	Not understandable	E7	0.38
7	Ignore user question	E2	0.37
2	General quality	E1	0.36
1	Unclear intention	E5	0.36
6	Ignore user utterance	E4	0.24
13	Word usage error	E3	0.18
16	Analysis failure	E1	0.17
4	Inappropriate answer	E4	0.17
3	Grammatical error	E3	0.15
12	Violation of common sense	E6	0.15
8	Contradiction	E2	0.14
5	Repetition	E2	0.11
9	Social error	E6	0.10
10	Diversion	E8	0.09
11	Topic-change error	E8	0.06
15	Mismatch in conversation	E7	0.06
14	Expression error	E3	0.02

Table 3: Correlation ratio (η) between the existence of a comment of a cluster (error type) and the number of breakdown labels. "Cat" denotes the error category of an error type.

Table 3 shows the correlation ratios for the error types. Clearly, not all error types have the same level of correlation. At the top of the table, there are four salient error types with similar η values: "Not understandable", "Ignore user question", "General quality", and "Unclear intention". Putting aside "General quality", which seems to concern the overall dialogue ability, the error types that we need to consider as fatal would seem to be the other three. Other errors seem to be less important with lower η values. "Expression error", which concerns the use of unnatural expressions, was found the least important.

When we look at the error categories, we can see an interesting result that it is NOT the error category that determines the fatality of errors but the specificity of error types. For example, "Not understandable" and "Mismatch in conversation" are both under error category E7 but have totally different effects on perceived breakdown. The same can be said for error types in E2.

Note that, although the values of correlation ratio seem rather low, the correlation often becomes low when it comes to subjective judgments (Higashinaka et al., 2004). Considering that we deal with chat-oriented dialogues, which are less restricted than task-oriented ones, we consider the current values of correlation ratio to be acceptable. Here, the important finding is that several error types are comparatively more important than the others.

4 Related work

Few studies have analyzed breakdowns in conversation. One exception is the study by Martinovsky and Traum (2003), who discussed possible causes of breakdowns they observed. Our work is different in that we systematically identify error types and quantitatively evaluate their effect. Our work can be seen as listing up errors in dialogue systems. A number of studies have aimed to create a taxonomy of errors (Bernsen et al., 1996; Möller, 2002; Paek, 2003), but their taxonomies are created manually and focus on task-oriented dialogue systems.

5 Summary and future work

By processing dialogue data with dialogue breakdown annotations and comments, this paper identified 17 error types that can be further categorized into eight error categories. By calculating correlation ratios, we discovered three error types that can be fatal: "Not understandable", "Ignore user question", and "Unclear intention". To avoid dialogue breakdowns, it is suggested that we need to make clear the meanings of system utterances, not ignore user questions, and show some intention behind system utterances. The findings will be useful for dialogue system developers who want to realize smooth human-machine interaction in chat-oriented dialogue systems and possibly in dialogue design as a whole.

For future work, we plan to consider ways to improve systems on the basis of our findings and also verify the generality of the results on data using other systems. To accurately detect dialogue breakdowns, dialogue systems researchers will need to collaborate. To this end, we are planning to organize an evaluation workshop on dialogue breakdown detection. For use in the evaluation workshop as well as in dialogue research in general, we have released our data with all the annotations and comments to the public.[1]

Acknowledgments

The "Project Next NLP" project in Japan is addressing the problems related to analyzing errors in natural language processing systems. One subgroup, comprising more than 32 researchers from 15 institutions, is collaborating in the performance of a dialogue task. We thank the members of this sub-group for the data collection, annotation, and fruitful discussions. We also thank NTT DOCOMO for letting us use their chat API for data collection.

References

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 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.

Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, and Kiyoaki Aikawa. 2004. Evaluating discourse understanding in spoken dialogue systems. *ACM Transactions on Speech and Language Processing (TSLP)*, 1:1–20.

Ryuichiro Higashinaka, Noriaki Kawamae, Kugatsu Sadamitsu, Yasuhiro Minami, Toyomi Meguro, Kohji Dohsaka, and Hirohito Inagaki. 2011. Unsupervised clustering of utterances using non-parametric bayesian methods. In *Proc. INTERSPEECH*, pages 2081–2084.

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.

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. 2002. A new taxonomy for the quality of telephone services based on spoken dialogue systems. In *Proc. SIGDIAL*, pages 142–153.

[1] https://sites.google.com/site/dialoguebreakdowndetection/

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.

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

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