On Dialogue Breakdown: Annotation and Detection

A report from the dialogue breakdown detection challenge

Kotaro Funakoshi[1], Ryuichiro Higashinaka[2], Michimasa Inaba[3], Yuka Kobayashi[4], Saku Sugawara[5], Katsuya Takanashi[6], Hiroko Otsuka[7], Hanae Koiso[8], and Mayumi Bono[9]

[1] Honda Research Institute Japan
[2] NTT Laboratories
[3] Hiroshima City University
[4] Toshiba Corporation
[5] University of Tokyo
[6] Kyoto University
[7] Hakodate Mirai University
[8] National Institute for Japanese Language and Linguistics
[9] National Institute of Informatics

Abstract. We organized the dialogue breakdown challenge, which is an evaluation workshop dedicated to the detection of breakdowns in Japanese human-machine chat-oriented dialogues. This paper reports analyses made on the breakdown annotation to the evaluation data used in the challenge and on behaviors of the detection systems submitted by six teams. The performances of the ensembles of the systems are also reported. Moreover, we investigate the effects of aggregating multiple models learned from different annotators. In response to the results of the investigation and in consideration of the highly subjective nature of breakdown, we propose '*not deciding unique correct labels*' in both training and evaluation of the systems.

1 Introduction

Dialogue breakdowns [5] in human-machine dialogues are caused for various reasons including errors and limited capabilities of machines, and are one of the major issues to be addressed in dialogue systems research. We collected Japanese dialogue breakdown data and proposed a taxonomy of errors by analyzing the data [3]. Providing the collected data as training data, we organized the dialogue breakdown detection challenge [4] and got six participating teams. The present paper shows the analysis results of the dialogue breakdown annotation on the evaluation data that was additionally collected for the challenge, and the analysis results of the behaviors of the submitted detection systems. The annotation of dialogue breakdown is highly subjective and thus exhibits large variances among annotators. We analyze the data and discuss the results with attention to such an aspect.

2 Data

We have two data sets to be investigated in this paper. One is the dialogues data that was annotated on breakdown by 30 persons per dialogue. The other is the results of 15 systems from six teams that tackled the detection challenge [4]. We call the former "the breakdown data" and the latter "the detection results."

2.1 The breakdown data

We collected 100 chat-oriented dialogues from anonymous workers in a crowdsourcing service.[10] Each worker accessed a designated web site though a browser and made a 20-turn conversation with a chat bot system. A conversation consists of 10 user utterances, 10 system responses, and one initial system prompt (such as "Hello! How are you?"). In the challenge, we provided 20 dialogues out of 100 as development data and reserved the rest 80 as evaluation data. Here we analyze this reserved evaluation data, which contains 800 system utterances (10 system utterances $\times$ 80 dialogues).

The system utterances in the breakdown data were annotated by anonymous workers in another crowdsourcing service.[11] For each dialogue, 30 workers annotated the system utterances on breakdown with the three labels of **NB**, **PB**, and **B**. Workers were different among the dialogues but each dialogue was annotated by a fixed group of 30s.

The definitions of **NB**, **PB**, and **B** were given to the annotators as follows.

NB (Not a Breakdown): It is easy to continue the conversation.
PB (Possible breakdown): It is difficult to continue the conversation smoothly.
B (Breakdown): It is difficult to continue the conversation.

While based on majority voting, we used a threshold t to control the rigor on **PB** and **B**. That is, only when a majority label got a vote rate higher than or equal to t, we assigned the label as the correct label. Otherwise, the correct label was set to **NB**.

In what follows, we set t to 0.5. Therefore, if the correct label of a system utterance is **PB** or **B**, it means more than 14 out of 30 persons chose the label for the utterance. In consequence, we had 473 **NB**s, 129 **PB**s, and 198 **B**s as the correct labels for the 800 system utterances.

2.2 The detection results

The six teams which joined the challenge submitted up to three results of their detectors' performances under different settings. We call each result "run", and we have 15 runs in total. Each run assigned one of the three labels of **NB**, **PB**, and **B** with confidences as probability distributions to all the 800 system utterances.

[10] http://crowdworks.jp/
[11] http://crowdsourcing.yahoo.co.jp/

3 Analysis on the breakdown data

3.1 Validation of majority vote annotation

We asked a helper (she is not an author of this paper, but a professional annotator working on various types of linguistic annotation and resource authoring) to check whether the labels decided by majority voting are subjectively acceptable for the helper given the same definition of breakdown.

Under the threshold $t = 0.5$, the helper reported sense of discomfort only with 32 utterances (4% of 800). Although this was just one person's introspection, we dare suppose that the annotation was mostly valid.

With regard to the 32 utterances in question, we found that the threshold was the cause of discomfort. For example, the following system utterance had label **NB**, although it posed an irrelevant question to the user's first question. The annotation details of 30 workers was 5, 14, and 11 for **NB**, **PB**, and **B** respectively. Due to the threshold setting, the least label (**NB**) was assigned as the correct label.

ID:1437630091
01 U: Konnichiwa, anata no namae wa?
(Hello, what is your name?)
02 S: Gokazoku towa issho ni kurasi te imasu ka?
(Are you living with your family together?)

Threshold t was introduced to exclude utterances that had low agreements among annotators from the evaluation, given the task was focusing on the detection of breakdown (i.e., **PB** and **B**). Therefore, we had to carefully pay attention to the threshold when we analyzed the data including **NB** and we used the data for training. For such purposes, it would be appropriate to decide the label by a pure majority vote without a threshold or to use the data while referring to the details of annotation (frequency distribution of labels). Indeed, as we will see later, it would be better not to use the uniquely fixed labels but to directly use frequencies of assigned labels for model training.

In short, although our definition of breakdown was very subjective and it produced considerable variations among annotators, we consider that our framework of breakdown data collection using the subjective definition and majority voting works for our purpose within a reasonable cost of corpus construction.

On the other hand, as shown by the above case of ID:1437630091, a finer inspection is needed. The above case seems to be a clear example of breakdown (**B**) at first glance, but in reality **PB** got the majority. In practice, given the above case, according to the definition of breakdown (**B**) "It is difficult to continue the conversation", there is no problem to continue the dialogue if one has no reluctance to abandon one's own preceding question. How to define the relation between 'breakdown' and 'deviation from conversational code' is an issue identified by our activity of the dialogue breakdown detection challenge.

Table 1. Frequency per error category

Main category	Sub-category	Worker A	Worker B
Utterance	Syntactic error	11	4
	Semantic error	33	18
	Uninterpretable	2	7
Response	Excess/lack of information	205	9
	Non-understanding	22	100
	No relevance	118	74
	Unclear intention	78	78
	Misunderstanding	4	0
Context	Excess/lack of proposition	41	35
	Contradiction	13	8
	Non-relevant topic	5	3
	Unclear relation	12	18
	Topic switch error	43	16
Environment	Lack of common ground	36	0
	Lack of common sense	11	0
	Lack of sociality	2	10

3.2 Types and distributions of errors

We categorized breakdowns (**PB** and **B**) in the 80 dialogues according to the error taxonomy in [3]. Only based on the descriptions of errors in a Japanese manuscript corresponding to [3], two workers (A and B) independently categorized breakdowns without any consensus building between them. We allowed them to categorize a breakdown to multiple categories so that we can know what types of errors are frequent. Table 1 shows the results of categorization.

With regards to the main categories of utterance and context, we see not so considerable discrepancies between workers A and B. In the main category of response, two workers showed considerably different behaviors on 'excess/lack of information' and 'non-understanding'. In the main category of environment, they categorized utterances differently, too. We may need to elaborate the definitions of these sub-categories.

It is necessary that annotators can categorize dialogue breakdowns with an adequate level of agreement and consistency so that we decompose breakdown detection into sub-problems and identify the technical bottlenecks. Therefore, we will have to improve both the taxonomy and categorization guidelines.

4 Analysis on the detection results

Table 2 shows the overview of six teams that participated in the challenge.[12] team1 submitted two results (run1 and run2, using RNN and LSTM-RNN respectively). Each of team2 to team 5 submitted three results while varying parameter settings. team6 submitted only one result. Thus we have 15 results

[12] More information is available in [4].

Table 2. Overview of teams

Team	Method	Used features
team1	RNN, LSTM-RNN	Word frequencies (bag-of-words), Cooccurrence frequencies of words between utterance-response pair, Sent2Vec encoding of word and cooccurrence frequencies.
team2	LSTM-RNN	Word2Vec encoding of word frequencies
team3	handcrafted rules	Keywords extracted by a Japanese morphological analyzer
team4	Poly kernel SVM	Word frequencies both in the target system utterance and in the previous user utterance,
team5	6-layer DNN	Estimated and predicted speech act types, perplexity of the target utterance, the estimated question classification results of the previous utterance.
team6	LSTM-RNN	Encoding of word frequencies by use of NCM, LSTM, bag-of-words embedding, an extended NCM
baseline	CRF	Word frequencies

(RNN:Recurrent Neural Network, LSTM:Long Short-Term Memory, DNN:Deep Neural Network, NCM:Neural Conversational Model)

Table 3. Number of utterances over the number of given **B** labels in the evaluation data

# of **B** labels given by annotators	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	total
# of such utterances	27	30	21	21	17	17	9	12	15	5	11	10	3	0	0	0	198

(More than 14 **B** labels are only shown as threshold t is 0.5)

(hereafter, systems) and denote these 15 systems as s1, ..., s15. When we want to mention a system in a more distinguishable way, we use the format of 'team$N$run$M$'. s1 is team1run1, s2 is team1run2, s3 is team2run1, and so on. The baseline system was provided by the organizers.

4.1 Correlation between the entropy of breakdown labels and systems' collective accuracy

The correct labels were decided by majority voting of 30 annotators. Unsurprisingly, the degrees of agreement were diverse. As Table 3 shows, 27 out of 198 utterances were labeled as **B** by 15 out of 30 annotators (50%), while only 3 out of 198 were labeled as **B** by 27 out of 30.

We can assume that a breakdown with a higher agreement is easier for human annotators to judge that it is a breakdown (and vice versa). Here we examine if this stands for systems, too. In concrete, we look into the correlation between the entropy of breakdowns and the collective accuracy of systems (i.e., how many systems out of 15 correctly answer for each case). If we can find a negative correlation, it indicates that difficulty is similar between humans and systems, as the lower the entropy is (i.e., small variation) the higher the accuracy is.

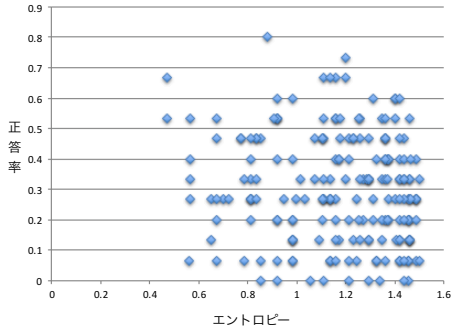

Fig. 1. Systems' collective accuracy versus entropy of breakdown labels

The calculated Pearson's correlation coefficient is -0.127. It is a very weak negative correlation although it does not contradict our hypothesis. As either visual observation cannot find any clear tendency in the plot shown in Figure 1, we must conclude that, even if an utterance is an obvious breakdown for human, it is not obvious for systems so far.

4.2 Differences in systems

Next we investigated if there was a similarity between systems, that is, whether systems answer correctly on the same set of data and answer wrongly on the rest of the data.

For this investigation, we counted how many systems answered correctly on each sample. As Table 4 shows, we had only 30 samples (15%) on which systems more than half answered correctly. In sum, the 15 systems were not so similar to each other. It seems too early yet to discuss the technical bottle necks, or what problems are really difficult, in breakdown detection.

4.3 Effectiveness of system aggregation

Because the systems behave differently, we may be able to build a better system by aggregating them. It is known that an aggregation or ensemble of machine learning-based components tend to produce better results in general by compensating 'unstableness' in individual components [1]. Henderson et al. also reported the best performance with the ensemble of the submitted systems in the dialogue state tracking challenge [2].

Table 4. Number of systems that correctly answered over the **B**-labeled utterances

# of systems that correctly answered	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
# of utterances	9	22	19	22	31	25	18	22	17	6	5	1	1	0	0	0	198

Table 5. Performances of ensemble systems on **B** detection

Aggregation method	Precision	Recall	F1
random baseline	.25 (66/267)	.33 (66/198)	.28
team5run2 (s13)	.33 (155/465)	.78 (155/198)	.47
OR(s1,...,s15)	.27 (189/707)	.96 (189/198)	.42
OR(s2,s3,s6,s15)	.43 (112/258)	.57 (112/198)	.49
VOTE(≥ 4)	.37 (126/338)	.64 (126/198)	.47
Classifier (SMO-RBF)	.47 (126/267)	.64 (126/198)	.54

Table 5 shows the results of the best single system and the three different aggregation methods: OR, VOTE, and Classifier. In F1 measure, Classifier performed the best.[13]

OR OR(s1,...,s15) represents a system that combines all the 15 systems with simple OR operation. That is, if any one of the 15 votes for **B**, the aggregated system votes for **B**. That naturally leads to higher recall but lower precision.

The best performance with OR operation was achieved when we combined only s2 (team1run2), s3 (team2run1), s6 (team3run1) and s15 (team5run1). This system showed lower recall than s13 but higher precision.

VOTE The VOTE($\geq N$) integrated system votes for **B**, when N systems out of the 15 vote for **B**. This system achieved the best performance when N was 4, but no considerable improvement was gained.

Classifier We can aggregate the systems by composing a classifier based on the 15 systems' output. Although it became almost equivalent to the common 'averaging' method if we adopted a simple linear classifier, here we used SMO (SVM impelementation in Weka data mining software) with RBF kernel and default parameters. The test was made by a 10-fold cross validation. Because the quantities of the labels are biased (**NB**: 473, **PB**: 129, **B**: 198), we used the weighting mechanism in Weka to remove the bias. In addition, we used not the predicted labels but confidences as features. Finally the aggregated system achieved the best performance in the F1 measure.

4.4 Annotator model aggregation

So far we used the correct labels that were reduced from 30 to one by majority voting. That means we built a collective virtual prototype personality from multiple persons' annotations, and tried to model the virtual personality by the use of a variety of methods. Here we consider to model multiple annotators individually by a single method, and then to aggregate the annotator models into one.

[13] Note that the result of Classifier was made by 10-fold cross validation, while the others were simply evaluated with the whole evaluation data.

Table 6. Aggregation of 24 individual models on **B** detection

	unified			aggregated			shuffled		
Aggregation method	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
RandomForest	.39	.37	.38	.44	.40	.42	.47	.46	.46
SMO (Poly-1)	.49	.34	.40	.46	.44	.45	.45	.42	.44
SMO (RBF)	.48	.27	.35	.45	.44	.45	.46	.48	.47

We trained 24 models from 24 person annotations in init100 data [3] with the baseline method using CRF shown in Table 2, then aggregated these 24 models with three machine learning methods.

Table 6 shows the results. Column group 'unified' means the performance of a model of a virtual personality made of 24 by majority voting. Group 'aggregated' means the performance of the aggregation of 24 individual annotator models. Apparently 'aggregated' shows better performances. This implies we have no active reason to fix a 'correct label' for each training sample, although we still have passive reasons such as training time and computational cost.

Here an interesting question is whether keeping the personality/consistency is the key factor to achieve better performance in our task settings or not. To test this question, we randomly shuffled the annotation data among 24 annotators at the unit of dialogue, and performed the same process as with 'aggregate'.

Group 'shuffled' in Table 6 shows the results of this test. The performance were almost equivalent between the aggregated and shuffled cases. This indicates that we can ask many different people to annotate the data and consequently we can diminish restrictions on annotation such as cost, period, platforms, etc.

In this experiment, after all, we evaluated the results with the 'uniquely fixed correct labels.' However, if we evaluate systems based on the similarities between distributions as we did in [4], we do not have to decide unique correct labels even for evaluations (i.e., we are released from picking a value of controversial and arbitrary threshold t). Although it is overhasty to conclude only from the results examined here, the approach of '*not deciding unique correct labels*' or multilabel/distribution-based training and evaluation, could be a standard for a highly subjective issue like breakdown detection.

References

1. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley-interscience (2001)
2. Henderson, M., Thomson, B., Williams, J.D.: The second dialog state tracking challenge. In: Proc. SIGDIAL. pp. 263–272 (2014)
3. Higashinaka, R., Funakoshi, K., Araki, M., Tsukahara, H., Kobayashi, Y., Mizukami, M.: Towards taxonomy of errors in chat-oriented dialogue systems. In: Proc. SIGDIAL. pp. 87–95 (2015)
4. Higashinaka, R., Funakoshi, K., Kobayashi, Y., Inaba, M.: The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In: Proc. LREC (2016)
5. Martinovsky, B., Traum, D.: The error is the cue: Breakdwon in human-machine interaction. In: Proc. Error Handling in Spoken Dialogue Systems. pp. 11–16 (2003)