Evaluating dialogue breakdown detection in chat-oriented dialogue systems

Yuiko Tsunomori[1], **Ryuichiro Higashinaka**[2], **Tetsuro Takahashi**[3], **Michimasa Inaba**[4]

[1]NTT DOCOMO, Inc., [2]NTT Media Intelligence Labs.
[3]Fujitsu Laboratories Ltd., [4]Hiroshima City University

yuiko.tsunomori.fc@nttdocomo.com

Abstract

The task of detecting dialogue breakdown, the aim of which is to detect whether a system utterance causes dialogue breakdown in a given dialogue context, has been actively investigated in recent years. However, it is not clear which evaluation metrics should be used to evaluate dialogue breakdown detectors, hindering progress in dialogue breakdown detection. We propose an approach of finding appropriate metrics for evaluating such detectors. In our approach, we first enumerate possible evaluation metrics then rank them on the basis of system ranking stability and discriminative power. By using the submitted runs (results of dialogue breakdown detection of participants) of a dialogue breakdown detection challenge, we experimentally found that MSE(NB+PB,B) and MSE(NB,PB,B), which represent the mean squared error calculated by comparing a detector's output distribution and a gold distribution, are appropriate metrics for dialogue breakdown detection.

1 Introduction

We are witnessing an emergence of chat-oriented dialogue systems due to their social and entertainment aspects (Wallace, 2009; Higashinaka et al., 2014a; Venkatesh et al., 2017). However, since the range of topics in chats is diverse and the content can be quite complex, it is often the case that such systems make inappropriate utterances, leading to dialogue breakdown, a situation in which a user and system cannot proceed with a dialogue (Martinovski and Traum, 2003).

To avoid such situations, the task of detecting dialogue breakdown (Higashinaka et al., 2016), the aim of which is to detect whether a system utterance causes dialogue breakdown in a given dialogue context, has been actively investigated. To cultivate techniques for detecting dialogue breakdown, evaluation workshops dedicated to dialogue breakdown detection (dialogue breakdown detection challenges; DBDCs) have been held (Higashinaka et al., 2017). The participants have developed dialogue breakdown detectors that output a dialogue breakdown label (B: breakdown, PB: possible breakdown, or NB: not a breakdown) and a probability distribution of these labels for a given system utterance in a given context. The definitions of the labels are given as follows.

NB: It is easy to continue the conversation after the system utterance in question.

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

B: It is difficult to continue the conversation after the system utterance in question.

For example, consider the following piece of dialogue.

S1: It's nice to go shopping alone.
U1: I agree. That's nice.
S2: Shopping takes time.
U2: Window shopping is also fun.
S3: It's fun to go shopping with somebody.

In the above dialogue, system utterances are prefixed with S and user utterances with U. The dialogue context is from S1 to U2, and the target utterance for dialogue breakdown detection is S3 (underlined). In this example, S3 is likely to cause a dialogue breakdown because S3 contradicts S1. Therefore, a detector that classifies this as B or PB will be regarded as accurate.

The performance of dialogue breakdown detectors in DBDCs is evaluated using a variety of evaluation metrics (there are nine metrics used in DBDCs), including those that compare a detector's output label with a gold label (classification-related metrics) and those that compare a detector's output distribution with a gold distribution (distribution-related metrics). The gold distribution is derived from the annotations of dialogue breakdown labels by multiple annotators; in DBDCs, 30 annotators have been used for each utterance to derive the gold distribution. Although several techniques for detecting dialogue breakdown have been proposed, the current problem is that, since there are many evaluation metrics used, it is not clear on which metric researchers should focus. To propel progress in dialogue breakdown detection, we should determine which metrics are appropriate.

We propose an approach of finding appropriate metrics for evaluating dialogue breakdown detectors. In our approach, we first enumerate possible evaluation metrics (22 in all), including those used in DBDCs as well as those we newly added. Then, we rank the evaluation metrics on the basis of two criteria, i.e., system ranking stability and discriminative power, that are used in information retrieval (IR) research (Webber et al.,). By using submitted DBDC runs (results of dialogue breakdown detection of participants), we experimentally found appropriate evaluation metrics.

In the next section, we cover related work. In Section 3, we describe our approach, including the enumeration of possible evaluation metrics and criteria for ranking the metrics. In Section 4, we present the ranking of the metrics and determine which are appropriate. Finally, in Section 5, we summarize the paper and mention future work.

2 Related work

There is a good body of work on detecting problematic situations in task-orientated dialogue systems (Walker et al., 2000b; Lendvai et al., 2002; Lopes et al., 2016; Meena et al., 2015). In these studies, features, such as speech-recognition results, language-understanding results, and prosodic information, were extracted from user/system utterances and used to train a model that can detect problematic situations (also called "miscommunications" or "hotspots").

Detecting problematic system utterances in chat-oriented dialogue systems has been actively studied. For example, Xiang et al. (2014) use machine-learning techniques to classify system utterances as problematic or non-problematic by using features related to user intent and user sentiment. Higashinaka et al. (2014b) proposed incorporating various dialogic features, such as dialogue-act types and question types, to detect incoherent system utterances. More recently, three series of DBDCs have been held (Higashinaka et al., 2017), and a number of teams participated and submitted their runs, showing growing interest in dialogue breakdown detection.

In contrast to this increasing attention, there has been little research on the evaluation metrics for dialogue breakdown detection. In past DBDCs, nine metrics were used without much emphasis on any one in particular, making it difficult for the participants to tune their detectors and for the organizers to determine the best detector. The problem is that, in task-oriented dialogue systems, problematic situations can be determined relatively easily with regards to the task at hand; however, in chat-oriented dialogue systems, deciding if an utterance is problematic can be highly subjective, making it difficult to define the gold label. The use of distribution-related metrics may solve this problem; however, it is not clear if they are any better than classification-related metrics.

In this study, we empirically verified which metrics are appropriate in dialogue breakdown detection. To this end, we turned to techniques used in IR studies and used the criteria of system ranking stability and discriminative power (see Section 3) to find appropriate evaluation metrics. Since IR-related work requires evaluating a system's output by comparing it with relevance assessment results obtained from multiple assessors, the setting of dialogue breakdown detection is similar to that in IR research; hence, the same technique can be applied. We acknowledge that the use of correlation is commonly used in dialogue research (Walker et al., 2000a; Higashinaka et al., 2004; Liu et al., 2016) to find appropriate evaluation metrics; however, this is only applicable when the target is a scalar value. In our case, gold data take the form of distributions, making the application of correlation-based approaches difficult.

A study on annotating chat-oriented dialogue systems with three labels (invalid, acceptable, valid) is currently underway in the WOCHAT initiative[1] (Charras et al., 2016; Curry and Rieser, 2016), but little research has been done to estimate these labels. Since the labels in that study are similar to those used in dialogue breakdown detection, we believe the proposed approach and the appropriate metrics found with the approach will be useful for that study.

3 Approach

We empirically verified which metrics are appropriate in dialogue breakdown detection. We first enumerated as many evaluation metrics as possible to create an exhaustive list of candidates for the metrics. Then, we ranked the metrics according to the selection criteria used in IR, i.e., system ranking stability and discriminative power.

3.1 Candidates for evaluation metrics

The metrics in DBDCs can be categorized into two types: classification-related and distribution-related (Higashinaka et al., 2016).

Classification-related metrics Classification-related metrics are used to evaluate the correctness of the classification of dialogue breakdown labels. These values are calculated by comparing the output label of the dialogue breakdown detector and the gold label determined by majority voting from the gold distribution. The value of a classification-related metric is calculated for each dialogue; for example, to derive an accuracy, we divide the number of correctly predicted labels by the total number of labels (system utterances) within a dialogue.

Distribution-related metrics Distribution-related metrics are used to evaluate the output probability distribution of dialogue breakdown labels, which are calculated by comparing the distribution of the labels predicted by the dialogue breakdown detector with the gold distribution. The value of a distribution-related metric is calculated for each utterance.

The nine evaluation metrics in past DBDCs are naturally our candidates. However, it is not clear whether these metrics are sufficient. Therefore, we added several evaluation metrics that we thought were worth considering. Table 1 lists all metrics used in this study; (2)–(6), (9)–(10), (14)–(16), and (20)–(22) are our newly added metrics.

We added (2) and (3) because, although cases in which PB+B or NB+PB is regarded as a single label were considered for mean squared error (MSE) and Jensen-Shannon divergence (JSD), these cases were not considered for accuracy. We also added (4)–(6), (9)–(10), (14)–(16), and (20)–(22), which are weighted metrics. Since we believe that utterances with a high agreement of annotations need to be treated with more emphasis than those with a low agreement, we devised weighted metrics. In this paper, we use the Simpson index for weighting. We calculate the weight w for each utterance with the following equation:

$$w = \sum_{l \in \{NB,PB,B\}} p_l^2, \tag{1}$$

[1] http://workshop.colips.org/wochat/

Table 1: Evaluation metrics. "+w" means that metrics are weighted. See Eq. (1) for deriving weight in weighted metrics.

	Metric	Description
		Classification-related metrics
(1)	Accuracy(NB,PB,B)	For the system utterances in a dialogue, we compare the predicted labels and their gold labels. Then, the accuracy is calculated by dividing the number of correctly classified labels by the total number of labels.
(2)	Accuracy(NB,PB+B)	Same as (1) when PB and B are regarded as a single label.
(3)	Accuracy(NB+PB,B)	Same as (1) when NB and PB are regarded as a single label.
(4)	Accuracy+w(NB,PB,B)	$c_n = \begin{cases} 1, & \text{if predicted label matches gold label;} \\ 0, & \text{otherwise;} \end{cases}$ $\text{Accuracy} = \frac{\sum_{n=1}^{N} c_n w_n}{\sum_{n=1}^{N} w_n}$ n means utterance index, N means the total number of utterances, and w means the weight.
(5)	Accuracy+w(NB,PB+B)	Same as (4) when PB and B are regarded as a single label.
(6)	Accuracy+w(NB+PB,B)	Same as (4) when NB and PB are regarded as a single label.
(7)	F1(B)	For the system utterances in a dialogue, we compare the predicted labels and their gold labels. Then, we derive the F1 for the classification of B labels by the harmonic mean of precision and recall for B labels. See (9) for the definition of precision and recall.
(8)	F1(PB+B)	Same as (7) when PB and B are regarded as a single label.
(9)	F1+w(B)	$pred_n(labels) = \begin{cases} 1, & \text{if predicted label is in } labels; \\ 0, & \text{otherwise;} \end{cases}$ $gold_n(labels) = \begin{cases} 1, & \text{if gold label is in } labels; \\ 0, & \text{otherwise;} \end{cases}$ $\text{TP} = \sum_{n=1}^{N} pred_n(\text{B}) gold_n(\text{B}) w_n$ $\text{FP} = \sum_{n=1}^{N} pred_n(\text{B}) gold_n(\text{NB, PB}) w_n$ $\text{TN} = \sum_{n=1}^{N} pred_n(\text{NB, PB}) gold_n(\text{NB, PB}) w_n$ $\text{FN} = \sum_{n=1}^{N} pred_n(\text{NB, PB}) gold_n(\text{B}) w_n$ $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ $\text{F1} = \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
(10)	F1+w(PB+B)	Same as (9) when PB and B are regarded as a single label.
		Distribution-related metrics
(11)	JSD(NB,PB,B)	For each system utterance, we compare the predicted distribution of the three labels (NB, PB, and B) and that of the gold labels. Then, Jensen-Shannon divergence is calculated.
(12)	JSD(NB,PB+B)	Same as (11) when PB and B are regarded as a single label.
(13)	JSD(NB+PB,B)	Same as (11) when NB and PB are regarded as a single label.
(14)	JSD+w(NB,PB,B)	The weighted version of (11). The value is weighted by w in Eq. (1).
(15)	JSD+w(NB,PB+B)	Same as (14) when PB and B are regarded as a single label.
(16)	JSD+w(NB+PB,B)	Same as (14) when NB and PB are regarded as a single label.
(17)	MSE(NB,PB,B)	For each system utterance, we compare the predicted distribution of the three labels (NB, PB, and B) and that of the gold labels. Then, mean squared error is calculated.
(18)	MSE(NB,PB+B)	Same as (17) when PB and B are regarded as a single label.
(19)	MSE(NB+PB,B)	Same as (17) when NB and PB are regarded as a single label.
(20)	MSE+w(NB,PB,B)	The weighted version of (17). The value is weighted by w in Eq. (1).
(21)	MSE+w(NB,PB+B)	Same as (20) when PB and B are regarded as a single label.
(22)	MSE+w(NB+PB,B)	Same as (20) when NB and PB are regarded as a single label.

where p_l means the probability of each label l in the gold probability distribution. For example, if the probability distribution is $(p_{NB}, p_{PB}, p_B) = (0.33, 0.33, 0.33)$, $w = 0.33$, and for $(p_{NB}, p_{PB}, p_B) = (0.0, 0.0, 1.0)$, $w = 1.0$. Thus, the higher the agreement of annotations is, the higher the weight of utterances becomes. In Table 1, weighted metrics are indicated with "+w." The use of this type of weighting has been considered in previous studies (Sakai, ; Shang et al., 2017) as "unanimity-aware gain" and has shown promising results, making systems more distinguishable; hence, our adoption of weighting.

3.2 Criteria of appropriate evaluation metrics

To select the most appropriate evaluation metrics from our metric candidates, we use two criteria (system ranking stability and discriminative power (Webber et al.,)) commonly used in IR. To calculate these values, we use the results of dialogue breakdown detection of multiple dialogue breakdown detection systems (typically called "runs" in evaluation workshops).

System ranking stability We can assume that an appropriate evaluation metric should be able to rank runs more or less in the same order independent of the dataset. System ranking stability can check whether the rankings of runs are stable across multiple datasets. To calculate stability, various datasets are prepared first. Then, for each dataset, the ranking of the runs is created. After that, the rank correlations of the ranking pairs are calculated and averaged to derive the system ranking stability.

Discriminative power We can assume that an appropriate evaluation metric should be as sensitive to the difference in runs as possible. By using each evaluation metric, we compare run pairs and see how many they significantly differ. We can regard the metrics with the most run pairs with statistically significant difference as the most appropriate evaluation metrics.

4 Evaluation

We experimentally searched for appropriate evaluation metrics that meet the criteria of system ranking stability and discriminative power. We ranked evaluation metrics for each language (note that the DBDC datasets contain both English and Japanese data) and calculated the average ranks so that we could select highly ranked ones across languages. In what follows, we describe the datasets we used and the procedure for calculating the values for the criteria.

4.1 Datasets

We used both the English and Japanese dialogue datasets of DBDC3[2] and the results of the submitted runs of the participants in DBDC3 (for details, see (Higashinaka et al., 2017)).

DBDC3 datasets The datasets were collected using four English systems [TKTK (Yu et al., 2016), IRIS (Banchs and Li, 2012), CIC[3], and YI[4]] and three Japanese systems [DCM (Onishi and Yoshimura, 2014), DIT (Tsukahara and Uchiumi, 2015), and IRS (IR-status-based system from (Ritter et al., 2011)]. Both datasets include 50 dialogue sessions, totaling 350 sessions. All dialogue sessions were 20 or 21 utterances long and included 10 system responses, each of which was annotated with dialogue breakdown labels by 30 annotators.

Submitted runs In the challenge, each participating team could submit up to three runs for each language. There were 12 runs for both English and Japanese. We also used the results of two baselines. One is a majority baseline that outputs the most frequent dialogue breakdown label in each system's development data with averaged probability distributions. The other was a baseline using conditional random fields (CRFs) that labels utterance sequences with the three breakdown labels by

[2] https://dbd-challenge.github.io/dbdc3/data/

[3] This dataset comes from the human evaluation round of the conversational intelligence challenge (http://convai.io/data/)

[4] https://www.slideshare.net/sld7700/skillbased-conversational-agent-80976302

Table 2: Submitted runs in English summarized by their key features. MemN2N and ETR denote end-to-end memory network and extra trees regressor, respectively.

Run	Model	Word/Sentence embedding	Bag of words	Utterance similarity	Turn index
KTH run1 (Lopes, 2017)	SVM			✓	
KTH run2	LSTM	✓			
KTH run3	LSTM	✓	✓		
PLECO run1 (Saito and Iki, 2017)	MemN2N	✓			
PLECO run2	MemN2N	✓			
RSL17BD run1 (Kato and Sakai, 2017)	ETR	✓		✓	✓
RSL17BD run2	ETR	✓		✓	✓
RSL17BD run3	ETR	✓		✓	✓
NCDS run1 (Park et al., 2017)	RNN	✓			
NCDS run2	RNN	✓			
NCDS run3	RNN	✓		✓	
SWPD run1 (Xie and Ling, 2017)	Bi-LSTM	✓			
CRF Baseline	CRF		✓		
Majority Baseline					

Table 3: Submitted runs in Japanese summarized by their key features. EoR denotes ensemble of regressors.

Run	Model	Word/Sentence embedding	Bag of words	Utterance similarity	Turn index
PLECO run1 (Saito and Iki, 2017)	MemN2N	✓			
PLECO run2	MemN2N	✓			
PLECO run3	MemN2N	✓			
RSL17BD run1 (Kato and Sakai, 2017)	ETR	✓		✓	✓
RSL17BD run2	ETR	✓		✓	✓
RSL17BD run3	ETR	✓		✓	✓
OUARS run1 (Takayama et al., 2017)	CNN	✓			
OUARS run2	CNN, LSTM	✓			
OUARS run3	CNN, LSTM	✓			
NTTCS run1 (Sugiyama, 2017)	EoR	✓		✓	✓
NTTCS run2	EoR	✓		✓	✓
NTTCS run3	EoR	✓		✓	✓
CRF Baseline	CRF		✓		
Majority Baseline					

using CRFs. The features used were words in a target utterance and the previous utterances. For the probability distribution, a probability of 1.0 was given to a label determined by the CRFs. Tables 2 and 3 summarize the submitted runs of the participants in English and Japanese, respectively. The tables indicate that many approaches have been tested, including those that use recent neural network models as well as those that use more conventional support vector machines (SVMs), random-forest-based methods such as extra trees regressor, and the ensemble of regressors.

4.2 Evaluation procedure

For system ranking stability, we used the rank correlation of ranked runs over different datasets to evaluate the metrics described in Section 3.2. There are two major rank-correlation statistics, Kendall's τ (Kendall, 1938) and Spearman rank correlation coefficient (Spearman, 1904). Because Kendall's τ has become a standard statistic for comparing the correlation between two ranked lists (Yilmaz et al., 2008), we used it to examine our rank correlation.

For both English and Japanese datasets, we first merged all data. Then, we created two subsets of data; each subset created by randomly sampling 20% from the merged data. For each metric, we ranked the runs for each subset to derive two run rankings. Finally, we calculated Kendall's τ between these rankings. To obtain stable results, we repeated this process 500 times and obtained the average value of Kendall's τ.

Regarding discriminative power, for each dataset of English and Japanese, we calculated the percentage of runs with statistical differences for all run pairs and ranked metrics according to that percentage. After that, we calculated the average rank over English and Japanese. We did this for each evaluation

Table 4: Results of system ranking stability

	English		Japanese		
Metrics	Kendall's τ	Rank	Kendall's τ	Rank	Average rank
MSE(NB+PB,B)	0.81	3	0.85	2	2.5
MSE(NB,PB,B)	0.79	6	0.86	1	3.5
MSE+w(NB+PB,B)	0.82	2	0.83	5	3.5
JSD(NB+PB,B)	0.81	4	0.83	4	4.0
JSD+w(NB+PB,B)	0.82	1	0.77	9	5.0
JSD(NB,PB,B)	0.77	12	0.85	3	7.5
JSD+w(NB,PB+B)	0.79	5	0.63	13	9.0
MSE(NB,PB+B)	0.78	11	0.77	8	9.5
JSD(NB,PB+B)	0.78	10	0.74	10	10.0
MSE+w(NB,PB+B)	0.78	8	0.68	12	10.0
MSE+w(NB,PB,B)	0.73	14	0.82	6	10.0
JSD+w(NB,PB,B)	0.75	13	0.78	7	10.0
Accuracy(NB+PB,B)	0.79	7	0.58	16	11.5
Accuracy+w(NB+PB,B)	0.78	9	0.61	15	12.0
Accuracy+w(NB,PB,B)	0.3	21	0.68	11	16.0
F1+w(B)	0.66	16	0.5	17	16.5
F1(B)	0.66	15	0.48	18	16.5
Accuracy(NB,PB,B)	0.26	22	0.63	14	18.0
F1(PB+B)	0.65	17	0.21	20	18.5
Accuracy(NB,PB+B)	0.62	18	0.18	21	19.5
Accuracy+w(NB,PB+B)	0.56	20	0.26	19	19.5
F1+w(PB+B)	0.61	19	0.14	22	20.5

Table 5: Average rank of each metric in terms of their discriminative power

	English		Japanese		
Metrics	% of run pairs with significant difference	Rank	% of pairs found with significant difference	Rank	Average rank
MSE(NB,PB,B)	67.0	6	76.9	2	4.0
MSE(NB+PB,B)	70.3	2	70.3	8	5.0
JSD(NB+PB,B)	67.0	6	74.7	4	5.0
MSE+w(NB+PB,B)	68.1	4	72.5	7	5.5
MSE(NB,PB+B)	68.1	4	64.8	9	6.5
MSE+w(NB,PB,B)	61.5	12	76.9	2	7.0
Accuracy(NB+PB,B)	71.4	1	52.7	14	7.5
JSD+w(NB+PB,B)	62.6	9	73.6	6	7.5
JSD(NB,PB,B)	60.4	14	81.3	1	7.5
JSD(NB,PB+B)	63.7	8	64.8	9	8.5
Accuracy+w(NB+PB,B)	70.3	2	50.5	16	9.0
JSD+w(NB,PB,B)	60.4	14	74.7	4	9.0
MSE+w(NB,PB+B)	61.5	12	60.4	11	11.5
F1+w(B)	62.6	9	50.5	16	12.5
F1(B)	62.6	9	48.4	18	13.5
JSD+w(NB,PB+B)	59.3	16	59.3	12	14.0
Accuracy+w(NB,PB,B)	19.8	21	58.2	13	17.0
Accuracy(NB,PB,B)	14.3	22	52.7	14	18.0
F1(PB+B)	56.0	17	15.4	22	19.5
Accuracy(NB,PB+B)	52.7	18	16.5	21	19.5
F1+w(PB+B)	50.5	19	17.6	20	19.5
Accuracy+w(NB,PB+B)	37.4	20	20.9	19	19.5

metric. We used `Discpower` [5] (Sakai, 2007) to calculate the discriminative power.

4.3 Results

Table 4 shows the ranking results for system ranking stability. Kendall's τ for both English and Japanese are shown. The average rank of the two ranks were used for the final measurement for stability. Overall, the distribution-related metrics (MSE, JSD) outperformed the classification-related ones. Among the distribution-related metrics, MSE(NB+PB,B) was the best in terms of system ranking stability. Also, the weighted metrics did not perform well when compared to the non-weighted ones, indicating that the weights were not that effective.

[5] `http://research.nii.ac.jp/ntcir/tools/discpower-en.html`

Table 6: Average rank of system ranking stability and discriminative power

Metrics	Rank of system ranking stability	Rank of discriminative power	Average rank
MSE(NB+PB,B)	2.5	5.0	3.8
MSE(NB,PB,B)	3.5	4.0	3.8
JSD(NB+PB,B)	4.0	5.0	4.5
MSE+w(NB+PB,B)	3.5	5.5	4.5
JSD+w(NB+PB,B)	5.0	7.5	6.3

Table 5 shows the results for discriminative-power evaluation (significance level $\alpha = .05$). We show the percentage of runs with statistically significant differences for all run pairs (the number of runs for both languages was 14; therefore, the number of all run pairs was $\binom{14}{2} = 91$). The distribution-related metrics (MSE, JSD) were ranked highly. Because the ranks of weighted metrics were low, similarly to the results for system ranking stability, our weighting did not seem to contribute much to discriminative power.

4.4 Determining appropriate metrics

Table 6 shows the top five evaluation metrics by their average rank for system ranking stability and discriminative power; MSE(NB+PB,B) and MSE(NB,PB,B) were the best evaluation metrics with the same average rank.

Because MSE and JSD were generally ranked high, we can confirm that the distribution-related metrics were more appropriate than the classification-related ones. This is probably because distribution-related metrics can use more information, which is lost when converting the distribution into a single label, as in classification-related metrics. We can also see that there was no difference between when NB and B were regarded as a single label, i.e., (NB+PB,B) and when all labels were separate, i.e., (NB,PB,B). Our speculation is that distinguishing between NB+PB and B is as difficult as distinguishing among the three labels. To verify this, we calculated the inter-annotator agreement (Fleiss' κ) of dialogue breakdown annotations. Regarding the English dataset, we found that when all labels are separate, κ is 0.065. When NB and PB are regarded as a single label, κ is 0.077, and when PB and B are regarded as a single label, κ is 0.095. The same tendency of κ was also found for the Japanese dataset. This indicates that distinguishing between NB+PB and B could be more difficult than between NB and PB+B and more similar to distinguishing among the three labels, supporting our speculation to some extent. In accordance with the results for system ranking stability and discriminative power, the weighted metrics were not effective. One possible reason could be that the weights are just making easy-to-guess problems stand out and de-emphasizing difficult-to-guess ones in the evaluation, making it difficult to differentiate the runs.

5 Summary and future work

To clarify which evaluation metrics should be used to evaluate dialogue breakdown detectors, we proposed an approach of finding the appropriate metrics for evaluating the detectors. We first enumerated possible evaluation metrics then ranked them on the basis of system ranking stability and discriminative power. By using the submitted runs, we experimentally found that MSE(NB+PB,B) and MSE(NB,PB,B) were appropriate metrics. As a final note, if we were to recommend a single metric, we suggest using MSE(NB+PB,B) because only two-way (NB+PB and B) annotations will be necessary, lowering the cost for preparing datasets.

For future work, we plan to consider combinations of multiple evaluation metrics to create more appropriate metrics. We also plan to enumerate other metrics because our list of metrics may not be sufficient. Although weight was found not to be that effective in this study, we plan to consider other weighting methods and pursue the reasons for their poor performance because we intuitively feel that weighting high-agreement utterances seems reasonable. Finally, we also want to improve the dialogue breakdown detector we are developing by using our proposed approach of finding evaluation metrics and improve our chat-oriented dialogue system.

References

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

Franck Charras, G Dubuisson Duplessis, Vincent Letard, Anne-Laure Ligozat, and Sophie Rosset. 2016. Comparing system-response retrieval models for open-domain and casual conversational agent. In *Proc. WOCHAT*.

Amanda Cercas Curry and Verena Rieser. 2016. A subjective evaluation of chatbot engines. In *Proc. WOCHAT*.

Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, and Kiyoaki Aikawa. 2004. Evaluating discourse understanding in spoken dialogue systems. *ACM Trans. Speech Lang. Process.*, 1:1–20, November.

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.

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

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. 2017. Overview of dialogue breakdown detection challenge 3. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Sosuke Kato and Tetsuya Sakai. 2017. RSL17BD at DBDC3: Computing utterance similarities based on term frequency and word embedding vectors. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Piroska Lendvai, Antal Van Den Bosch, Emiel Krahmer, and Marc Swerts. 2002. Improving machine-learned detection of miscommunications in human-machine dialogues through informed data splitting. In *Proc. the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics*, pages 1–15.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

José Lopes, Arodami Chorianopoulou, Elisavet Palogiannidi, Helena Moniz, Alberto Abad, Katerina Louka, Elias Iosif, and Alexandros Potamianos. 2016. The SpeDial datasets: datasets for spoken dialogue systems analytics. In *Proc. LREC*, pages 104–110.

José Lopes. 2017. How generic can dialogue breakdown detection be? the KTH entry to DBDC3. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Bilyana Martinovski and David Traum. 2003. Breakdown in human-machine interaction: the error is the clue. In *Proc. ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16.

Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. Automatic detection of miscommunication in spoken dialogue systems. In *Proc. SIGDIAL*, pages 354–363.

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

Chanyoung Park, Kyungduk Kim, and Songkuk Kim. 2017. Attention-based dialog embedding for dialog breakdown detection. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

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

Atsushi Saito and Taichi Iki. 2017. End-to-end character-level dialogue breakdown detection with external memory models. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Tetsuya Sakai. The effect of inter-assessor disagreement on IR system evaluation: A case study with lancers and students. *Proc. EVIA*, pages 31–38.

Tetsuya Sakai. 2007. Evaluating information retrieval metrics based on bootstrap hypothesis tests. *IPSJ Digital Courier*, 3:625–642.

Lifeng Shang, Tetsuya Sakai, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao, Yuki Arase, and Masako Nomoto. 2017. Overview of the NTCIR-13 short text conversation task. In *Proc. NTCIR*, pages 194–210.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Hiroaki Sugiyama. 2017. Dialogue breakdown detection based on estimating appropriateness of topic transition. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Junya Takayama, Eriko Nomoto, and Yuki Arase. 2017. Dialogue breakdown detection considering annotation biases. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

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

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2017. On evaluating and comparing conversational agents. *Proc. NIPS 2017 Conversational AI workshop*.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000a. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3-4):363–377.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You? In *Proc. NAACL*, pages 210–217.

Richard S Wallace. 2009. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

William Webber, Alistair Moffat, and Justin Zobel. The effect of pooling and evaluation depth on metric stability. In *Proc. EVIA*, pages 7–15.

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.

Zeying Xie and Guang Ling. 2017. Dialogue breakdown detection using hierarchical bi-directional LSTMs. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proc. the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander I Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proc. SIGDIAL*, pages 404–412.