Overview of Dialogue Breakdown Detection Challenge 3

Ryuichiro Higashinaka[1], *Kotaro Funakoshi*[2], *Michimasa Inaba*[3]
Yuiko Tsunomori[4], *Tetsuro Takahashi*[5], *Nobuhiro Kaji*[6]

[1]NTT Media Intelligence Labs.
[2]Kyoto University/Honda Research Institute Japan
[3]Hiroshima City University
[4]NTT DOCOMO, Inc.
[5]Fujitsu Laboratories Ltd.
[6]Yahoo Japan Corporation

dbdc3-organizers@googlegroups.com

Abstract

Dialogue breakdown detection is a promising technique for dialogue systems. To promote the research and development of such a technique, we have been organizing a dialogue breakdown detection challenge where the task is to detect a system's inappropriate utterances that lead to dialogue breakdowns in chat-oriented dialogue. In this paper, we present an overview of dialogue breakdown detection challenge 3. Unlike the previous two challenges held in Japan, we used datasets in two languages: English and Japanese. Overall, eight teams participated in the challenge, in which six teams worked on English and four teams on Japanese, with two teams working on both. We describe the task setting, datasets, and evaluation metrics for the challenge as well as the results of the submitted runs of the participants.
Index Terms: dialogue breakdown detection, chat-oriented dialogue systems, evaluation workshop

1. Introduction

In this paper, we present an overview of dialogue breakdown detection challenge 3 (DBDC3), which is the third evaluation workshop dedicated to dialogue breakdown detection [1]. The aim of the challenge was to bring together various methods for detecting dialogue breakdown so that the capability of chat-oriented dialogue systems can be improved.

The task of detecting dialogue breakdown is to detect whether a system utterance causes a dialogue breakdown (a situation in a dialogue where users cannot proceed with a conversation [2, 3, 4]) in a given dialogue context. The participants of the challenge develop a dialogue breakdown detector that outputs a dialogue breakdown label (B: breakdown, PB: possible breakdown, or NB: not a breakdown) and a distribution of these labels. The definition of the labels are defined 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.

Unlike the previous two challenges, in this challenge, we used datasets in two languages: English and Japanese. Overall, eight teams participated in the challenge, in which six teams worked on English and four teams on Japanese with two teams working on both. In this paper, we describe the task setting, datasets, and evaluation metrics for the challenge as well as the results of the submitted runs of the participants.

2. Task setting

The task of the dialogue breakdown detection challenge is to detect whether a system utterance causes a dialogue to break down in a given dialogue context. 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.

Here, system utterances are prefixed by S and user utterances by 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 be causing a dialogue breakdown because S3 contradicts S1. Therefore, a detector that classifies this as a dialogue breakdown will be regarded as accurate. The reason we say "likely" is that a human's decision on dialogue breakdown is highly subjective, and it is difficult to decide on a single gold label. For this reason, we use many annotators for dialogue breakdown annotation and opt for majority voting and their probability distribution as references.

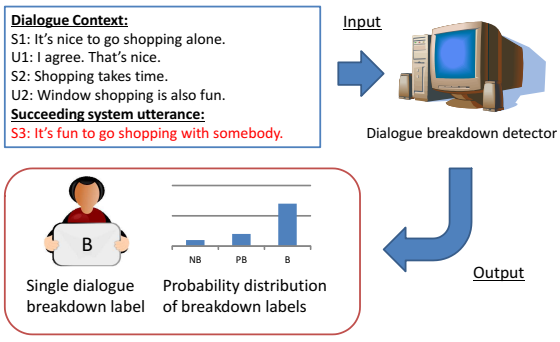

Figure 1: *Illustration of task setting*

Given pairs of dialogue context and a succeeding system utterance, the participants of the challenge submit, for each pair, (1) a single dialogue breakdown label and (2) the probability distribution of breakdown labels (See Fig. 1. Note that, although some utterances may exist after the target utterance, they

cannot be used for prediction because, for this challenge, we focus on avoiding dialogue breakdown rather than recovery. In the challenge, each participant can submit up to three "runs" for each language, so several parameters for dialogue breakdown detection can be tested.

3. Datasets

We distributed two sets of data to participants: one consisting of training data and the other of development and evaluation (test) data. Tables 1 and 2 show the statistics of the datasets for English and Japanese, respectively. In this section, we first describe the English datasets and then the Japanese ones.

3.1. Datasets for English

3.1.1. Development data

We provided four datasets: TKTK-100, IRIS-100, CIC-115, and YI-100. The dialogue data for TKTK-100 and IRIS-100 were taken from the WOCHAT TickTock and IRIS datasets[1]. The source of CIC-115 is the human evaluation round of the Conversational Intelligence Challenge (CIC)[2]. For YI-100, we newly collected dialogue sessions by crowdsourcing. All dialogue sessions were 20 or 21 utterances long and included 10 system responses. The four datasets are described in the following.

TKTK-100 We selected 100 sessions out of 206 in the original WOCHAT dataset. TickTock sessions start from user utterances. See [5] for the details of TickTock. Dialogue breakdown annotation was done by using a crowdsourcing service, CrowdFlower[3]. For each utterance, 30 different workers annotated each dialogue session. Level-2 workers[4] from Australia, Canada, New Zealand, UK, and USA were recruited, and we requested non-native English speakers to refrain from participating in our annotation tasks.

IRIS-100 We selected 100 sessions out of 163 in the original WOCHAT dataset. The dataset was processed in the same manner as TKTK-100. Original IRIS sessions start from system utterances; however, we cut the first system utterances to make the data format identical to that of TKTK-100 for annotation purposes. See [6] for the details of IRIS.

CIC-115 This dataset comes from the human evaluation round of the conversational intelligence challenge and DeepHack Turing school-hackathon[5]. The dialogues are available at the CIC site[6]. From all $2,778$ dialogue sessions, we selected 115 dialogues performed between a human and a bot. Within the 115 dialogues, 85 dialogues start with a system utterance, and 30 dialogues start with a user utterance. As per the convention of CIC, each dialogue comes with a short paragraph, which is used as the context of the dialogue. The paragraphs are from the SQuAD dataset[7]. Dialogue breakdown annotation was done by using a crowd-sourcing service, Amazon Mechanical Turk (AMT)[8], with 30 workers. When recruiting the workers, we specified that the task requires native English skills for the task instructions.

YI-100 We collected 100 dialogue sessions by using a chatbot developed at the Moscow Institute of Physics and Technology[9] by using AMT. A worker was assigned to have a chat with the system that was more than 10 utterances. Dialogue breakdown annotation was also done by using AMT with 30 workers. YI sessions start from system utterances.

3.1.2. Evaluation data

In the formal run, we distributed the following evaluation data.

TKTK-50 We collected 50 dialogue sessions of TickTock by using AMT in the same way as YI-100. Dialogue breakdown annotation was done by using CrowdFlower.

IRIS-50 We selected 50 dialogue sessions from the held-out IRIS data graciously provided by the IRIS team. Dialogue breakdown annotation was done by using CrowdFlower.

CIC-50 The data source of CIC-50 is held-out dialogue data collected by CIC after the human evaluation round. Dialogue breakdown annotation was done by using AMT.

YI-50 We collected 50 dialogue sessions of YI in the same way as YI-100. Dialogue breakdown annotation was done by using AMT.

3.2. Datasets for Japanese

As for the Japanese datasets, we did not create new development data because we had already created several datasets in the previous two challenges (DBDC1 and DBDC2). Here, we briefly describe the development data and the newly created evaluation data.

3.2.1. Development data

Chat dialogue corpus This dataset has 1,146 dialogue sessions. The dialogues were collected by using NTT DOCOMO's chat API (DCM) [7]. 100 dialogues (called init100) were annotated by 24 annotators, and the rest of the dialogues (called rest1046) were annotated by 2-3 annotators. Dialogue breakdown annotation was done by the researchers working on chat-oriented dialogue systems in Japan.

Development data for DBDC1 This dataset has 20 dialogue sessions. The dialogues were collected by using DCM. The dialogues were collected by using a crowd-sourcing service, CrowdWorks[10], and were annotated by 30 annotators by using another crowd-sourcing service, Yahoo! Crowd-sourcing[11]. All datasets in DBDC1 and DBDC2 were collected and annotated in the same way.

Evaluation data for DBDC1 This dataset contains 80 dialogue sessions. The dialogues are collected by using DCM.

[1] http://workshop.colips.org/wochat/data/index.html

[2] http://convai.io

[3] https://www.crowdflower.com/

[4] Higher quality: smaller group of more experienced, higher accuracy contributors by the definition of CrowdFlower.

[5] http://turing.tilda.ws

[6] http://convai.io/data/

[7] https://rajpurkar.github.io/SQuAD-explorer/

[8] https://requester.mturk.com

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

[10] http://crowdworks.jp

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

Table 1: *Statistics of English datasets*

	Development data				Evaluation data			
	TKTK-100	IRIS-100	CIC-115	YI-100	TKTK-50	IRIS-50	CIC-50	YI-50
No. of sessions	100	100	115	100	50	50	50	50
No. of annotators	30	30	30	30	30	30	30	30
NB (Not a Breakdown)	35.1%	32.9%	28.9%	34.8%	44.3%	34.5%	29.1%	35.4%
PB (Possible Breakdown)	27.6%	27.8%	29.8%	36.1%	29.2%	29.3%	39.3%	40.3%
B (Breakdown)	37.3%	39.4%	41.3%	29.1%	26.5%	36.2%	31.6%	24.3%
Fleiss' κ (NB, PB, B)	0.14	0.11	0.054	0.011	0.13	0.090	0.0040	−0.0060
Fleiss' κ (NB, PB+B)	0.21	0.15	0.084	0.020	0.19	0.13	0.0072	−0.0043

Table 2: *Statistics of Japanese datasets*

	Chat dialogue corpus		DBDC1	DBDC2 (DVL/EVL)			DBDC3 (EVL)		
	init100	rest1046	DVL/EVL	DCM	DIT	IRS	DCM	DIT	IRS
No. of sessions	100	1046	20/80	50/50	50/50	50/50	50	50	50
No. of annotators	24	2 or 3	30	30	30	30	30	30	30
NB (Not a Breakdown)	59.2%	58.3%	37.1%	39.8%	33.0%	37.4%	34.9%	25.3%	29.3%
PB (Possible Breakdown)	22.2%	25.3%	32.2%	30.2%	27.4%	24.3%	34.2%	28.3%	23.8%
B (Breakdown)	18.6%	16.4%	30.6%	29.9%	39.5%	38.3%	30.9%	46.4%	46.9%
Fleiss' κ (NB, PB, B)	0.28	0.28	0.20	0.31	0.24	0.36	0.24	0.14	0.27
Fleiss' κ (NB, PB+B)	0.40	0.40	0.27	0.44	0.38	0.48	0.32	0.20	0.37

Development data for DBDC2 This dataset has 150 dialogue sessions. The dialogues are collected by using DCM, DIT (Denso IT Laboratories' system) [8], and IRS (IR-status based system from [9]) systems.

Evaluation data for DBDC2 This dataset has 150 dialogue sessions, 50 dialogues each from DCM, DIT, and IRS.

3.2.2. Evaluation data

The evaluation data for Japanese contained 150 dialogue sessions. We used the same procedure that we used to create the evaluation data for DBDC2.

4. Evaluation metrics

As in the previous challenges, we used two types of evaluation metrics: classification-related and distribution-related.

4.1. Classification-related metrics

Classification-related metrics were used to evaluate the accuracy in classifying breakdown labels. Here, the accuracy is calculated by comparing the output of the detector and the gold label determined by majority voting. We use a threshold t to obtain the gold label; that is, we first find the majority label and check if the ratio of that label is above t. If so, the gold label becomes that label and NB otherwise. We used the following metrics.

- Accuracy: the number of correctly classified labels divided by the total number of labels to be classified.
- Precision, Recall, F-measure (B): the precision, recall, and F-measure for the classification of B labels.
- Precision, Recall, F-measure (PB+B): The precision, recall, and F-measure for the classification of PB + B labels; that is, PB and B labels are treated as a single label.

These metrics can provide intuitive results about the detection of dialogue breakdowns because they are used to directly evaluate whether dialogue breakdowns are correctly classified or not. However, the choice of an appropriate t value remains an open issue. In this challenge, we used $t = 0.0$, which means simple majority voting.

4.2. Distribution-related metrics

Distribution-related metrics were used to evaluate the similarity of the distribution of breakdown labels, which is calculated by comparing the predicted distribution of the labels with that of the gold labels. We calculate these values for each utterance and use the mean values for evaluation. We used the following metrics.

- JS Divergence (NB,PB,B): distance between the predicted distribution of the three labels and that of the gold labels calculated by Jensen-Shannon Divergence.
- JS Divergence (NB,PB+B): JS divergence when PB and B are regarded as a single label.
- JS Divergence (NB+PB,B): JS divergence when NB and PB are regarded as a single label.
- Mean Squared Error (NB,PB,B): distance between the predicted distribution of the three labels and that of the gold labels calculated by mean squared error.
- Mean Squared Error (NB,PB+B): mean squared error when PB and B are regarded as a single label.
- Mean Squared Error (NB+PB,B): mean squared error when NB and PB are regarded as a single label.

These metrics are used to compare the distributions of the labels, thus enabling a direct comparison with the gold labels However, the results may not be as easily interpretable as the classification-related metrics because they do not directly translate to detection performance.

Table 3: *Description of systems for English. All teams opted in to disclose their team/organization names. An asterisk indicates an unofficial run without technical paper submission.*

Team	Organization	# of runs	Method	Description of runs
SAM2017	Soochow University	1	Maximum entropy model	***run1:** To solve this problem, we split a conversation into a system-to-user conversation (U-S).Thus, a file contained 10 pairs of (U-S), and we got the label for each system statement (S:) by using a classifier trained by the maximum entropy model.
KTH	KTH	3	LSTM, SVM	**run1:** SpeDial project reduced feature set + SVM; **run2:** pre-trained word embeddings + LSTM; **run3:** BoW + document embeddings + LSTM.
PLECO	Nextremer	2	MemNN	**run1:** any utterance of a speaker is given as a sequence of characters. Our end-to-end breakdown detection model is trained by plain SGD to minimize softmax-cross-entropy. The model consists of a plain memory network with attention over attention; **run2:** run1 + training data consists of multilingual dialogue breakdown training data.
RSL17BD	Waseda University	3	ETR	**run1:** The calculation of utterance similarities uses maximum cosine similarities between terms and the geometric mean; **run2:** run1 + the calculation of utterance similarities uses the arithmetic mean of cosine similarities between terms; **run3:** run1 + the calculation of utterance similarities uses maximum cosine similarities between terms and the arithmetic mean.
NCDS	Naver	3	RNN	**run1:** RNN + attention (between sentences) model; **run2:** same model as run1 but with different hyper-parameters; **run3:** some manually calculated feature vectors concatenated at the last layer.
SWPD	Baidu	1	Bi-LSTM	**run1:** word representation: pre-trained GloVe word embeddings, utterance representation: computed by a Bi-LSTM over words of a utterance, every system utterance embedding is concatenated with the previous user utterance embedding, output representation: computed by another Bi-LSTM over the utterances of a dialogue.

Table 4: *Description of systems for Japanese. All teams opted in to disclose their team/organization names.*

Team	Organization	# runs	Method	Description of runs
PLECO	Nextremer	3	MemNN	**run1:** any utterance of a speaker is given as a sequence of characters. Our end-to-end breakdown detection model is trained by plain SGD to minimize softmax-cross-entropy. The model consists of a plain memory network with attention over attention; **run2:** run1 + training to learn the sentence prediction sub-task with an extra training data set; **run3:** run1 + training data consists of multilingual dialogue breakdown training data.
RSL17BD	Waseda University	3	ETR	**run1:** the calculation of utterance similarities uses maximum cosine similarities between terms and the geometric mean; **run2:** run1 + the calculation of utterance similarities uses the arithmetic mean of cosine similarities between terms; **run3:** run1 + the calculation of utterance similarities uses maximum cosine similarities between terms and the arithmetic mean.
OUARS	Osaka University	3	LSTM, CNN	**run1:** clustering of train data based on the annotation distribution of each annotator. Construction of classifiers using parallel CNN encoder with each cluster. An ensemble of all classifiers; **run2:** run1 + using series LSTM in place of parallel CNN; **run3:** An ensemble of run1 and run2.
NTTCS	NTT Communication Science Laboratories	3	Ensemble of regressors	**run1:** Ensemble of regressors, without t-SNE reduction, DBDC1+2; **run2:** run1 + with t-SNE reduction, DBDC1+2; **run3:** run1 + without t-SNE reduction, DBDC2.

5. Results

Overall, eight teams participated in the challenge, in which six teams worked on English and four teams on Japanese, with two teams working on both. Each team could submit up to three runs for each language. We had 13 runs and 12 runs for English and Japanese, respectively.

Tables 3 and 4 show descriptions of the submitted runs of the participants in English and Japanese, respectively. We also show the results of two baselines. One is a majority baseline that outputs the most frequent dialogue breakdown label in each system's development data for English and development and evaluation data of DBDC2 for Japanese with averaged prob-

Table 5: *Overall Results of Classification (English)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
KTH run2	0.4415	PLECO run1	0.3636	Majority Baseline	0.8927
RSL17BD run2	0.4310	PLECO run2	0.3565	PLECO run1	0.8744
SWPD run1	0.4295	CRF Baseline	0.3543	PLECO run2	0.8708
CRF Baseline	0.4285	KTH run1	0.3487	KTH run1	0.8423
RSL17BD run1	0.4265	KTH run3	0.3373	RSL17BD run2	0.8400
KTH run3	0.4220	Majority Baseline	0.3343	RSL17BD run3	0.8357
RSL17BD run3	0.4200	SWPD run1	0.3210	RSL17BD run1	0.8329
SAM2017 run1	0.4060	RSL17BD run2	0.3201	NCDS run3	0.8046
Majority Baseline	0.3720	NCDS run3	0.3198	SWPD run1	0.7627
NCDS run2	0.3655	RSL17BD run1	0.3126	CRF Baseline	0.7622
NCDS run1	0.3605	RSL17BD run3	0.3025	KTH run3	0.7592
NCDS run3	0.3565	KTH run2	0.2949	KTH run2	0.7440
KTH run1	0.3375	SAM2017 run1	0.2413	NCDS run1	0.3458
PLECO run1	0.2950	NCDS run2	0.2097	NCDS run2	0.3397
PLECO run2	0.2900	NCDS run1	0.2076	SAM2017 run1	0.2160

Table 6: *Overall Results of JS Divergence (English)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
Majority Baseline	0.0393	Majority Baseline	0.0237	RSL17BD run2	0.0225
NCDS run1	0.0412	NCDS run1	0.0248	RSL17BD run3	0.0243
RSL17BD run2	0.0412	NCDS run2	0.0248	RSL17BD run1	0.0247
NCDS run2	0.0412	RSL17BD run2	0.0256	NCDS run2	0.0254
RSL17BD run3	0.0426	RSL17BD run3	0.0258	NCDS run1	0.0254
RSL17BD run1	0.0432	RSL17BD run1	0.0263	Majority Baseline	0.0257
KTH run2	0.0481	KTH run2	0.0267	KTH run2	0.0262
NCDS run3	0.0668	PLECO run1	0.0427	SWPD run1	0.0444
PLECO run1	0.0714	NCDS run3	0.0436	NCDS run3	0.0488
PLECO run2	0.0774	SWPD run1	0.0438	PLECO run1	0.0535
SWPD run1	0.0807	PLECO run2	0.0482	PLECO run2	0.0565
SAM2017 run1	0.2823	KTH run3	0.1892	SAM2017 run1	0.0805
KTH run3	0.3268	KTH run1	0.2343	KTH run1	0.2058
CRF Baseline	0.4409	SAM2017 run1	0.2377	KTH run3	0.2166
KTH run1	0.4445	CRF Baseline	0.2687	CRF Baseline	0.2985

Table 7: *Overall Results of Mean Squared Error (English)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	MSE	MSE (NB+PB,B)
Majority Baseline	0.0224	Majority Baseline	0.0278	RSL17BD run2	0.0246
NCDS run1	0.0237	NCDS run1	0.0287	Majority Baseline	0.0264
NCDS run2	0.0237	NCDS run2	0.0288	NCDS run2	0.0270
RSL17BD run2	0.0241	RSL17BD run2	0.0301	NCDS run1	0.0270
RSL17BD run3	0.0250	RSL17BD run3	0.0301	RSL17BD run3	0.0271
RSL17BD run1	0.0254	RSL17BD run1	0.0307	RSL17BD run1	0.0275
KTH run2	0.0281	KTH run2	0.0315	KTH run2	0.0286
PLECO run1	0.0415	PLECO run1	0.0455	SWPD run1	0.0497
NCDS run3	0.0437	SWPD run1	0.0501	NCDS run3	0.0572
PLECO run2	0.0448	PLECO run2	0.0509	SAM2017 run1	0.0621
SWPD run1	0.0471	NCDS run3	0.0677	PLECO run1	0.0632
SAM2017 run1	0.1441	KTH run3	0.1664	PLECO run2	0.0673
KTH run3	0.1670	KTH run1	0.1752	KTH run1	0.1476
CRF Baseline	0.2185	CRF Baseline	0.2171	KTH run3	0.2044
KTH run1	0.2240	SAM2017 run1	0.2652	CRF Baseline	0.2578

ability distributions. The other is a conditional random field (CRF) based baseline that labels utterance sequences with the three breakdown labels by using CRFs. The features used were words in a target utterance and its previous utterances. As for the probability distribution, the probability of 1.0 is given to the label determined by the CRFs.

Table 8: *Overall Results of Classification (Japanese)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
NTTCS run1	0.6129	NTTCS run1	0.6714	RSL17BD run2	0.8297
NTTCS run2	0.6085	NTTCS run2	0.6684	OUARS run3	0.8214
NTTCS run3	0.6017	NTTCS run3	0.6641	RSL17BD run1	0.8203
OUARS run3	0.5669	OUARS run3	0.6364	PLECO run2	0.8144
OUARS run1	0.5613	OUARS run1	0.6316	RSL17BD run3	0.8137
OUARS run2	0.5514	OUARS run2	0.6230	PLECO run1	0.8136
PLECO run3	0.5386	PLECO run3	0.6193	OUARS run2	0.8086
CRF Baseline	0.5322	PLECO run2	0.6087	NTTCS run3	0.8055
PLECO run1	0.5146	PLECO run1	0.6072	OUARS run1	0.8022
PLECO run2	0.5067	CRF Baseline	0.5796	PLECO run3	0.7971
Majority Baseline	0.4791	Majority Baseline	0.4511	NTTCS run1	0.7918
RSL17BD run3	0.4049	RSL17BD run3	0.2844	NTTCS run2	0.7909
RSL17BD run2	0.3924	RSL17BD run2	0.2795	CRF Baseline	0.7774
RSL17BD run1	0.3900	RSL17BD run1	0.2635	Majority Baseline	0.5580

Table 9: *Overall Results of JS Divergence (Japanese)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
NTTCS run1	0.0691	NTTCS run1	0.0466	NTTCS run1	0.0423
NTTCS run2	0.0693	NTTCS run2	0.0468	NTTCS run2	0.0424
NTTCS run3	0.0719	NTTCS run3	0.0489	NTTCS run3	0.0442
OUARS run1	0.0891	OUARS run1	0.0617	OUARS run1	0.0551
OUARS run3	0.0912	OUARS run3	0.0639	OUARS run3	0.0563
PLECO run3	0.0959	PLECO run3	0.0679	OUARS run2	0.0596
OUARS run2	0.0968	OUARS run2	0.0685	PLECO run3	0.0601
PLECO run2	0.0985	PLECO run2	0.0698	PLECO run2	0.0616
PLECO run1	0.1121	RSL17BD run2	0.0714	PLECO run1	0.0727
Majority Baseline	0.1299	RSL17BD run3	0.0751	Majority Baseline	0.0750
RSL17BD run2	0.1528	RSL17BD run1	0.0761	RSL17BD run1	0.0964
RSL17BD run1	0.1539	PLECO run1	0.0807	RSL17BD run2	0.0967
RSL17BD run3	0.1543	Majority Baseline	0.1058	RSL17BD run3	0.0980
CRF Baseline	0.3851	CRF Baseline	0.2379	CRF Baseline	0.2797

Table 10: *Overall Results of Mean Squared Error (Japanese)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
NTTCS run1	0.0372	NTTCS run1	0.0454	NTTCS run1	0.0469
NTTCS run2	0.0373	NTTCS run2	0.0456	NTTCS run2	0.0470
NTTCS run3	0.0388	NTTCS run3	0.0482	NTTCS run3	0.0487
OUARS run1	0.0465	OUARS run1	0.0582	OUARS run1	0.0581
OUARS run3	0.0469	OUARS run3	0.0593	OUARS run3	0.0583
OUARS run2	0.0498	OUARS run2	0.0638	OUARS run2	0.0615
PLECO run3	0.0517	PLECO run3	0.0667	PLECO run3	0.0655
PLECO run2	0.0543	PLECO run2	0.0715	PLECO run2	0.0678
PLECO run1	0.0627	RSL17BD run2	0.0737	Majority Baseline	0.0750
Majority Baseline	0.0676	RSL17BD run3	0.0780	PLECO run1	0.0813
RSL17BD run2	0.0879	RSL17BD run1	0.0785	RSL17BD run1	0.1033
RSL17BD run1	0.0882	PLECO run1	0.0829	RSL17BD run2	0.1040
RSL17BD run3	0.0886	Majority Baseline	0.1012	RSL17BD run3	0.1047
CRF Baseline	0.1996	CRF Baseline	0.2081	CRF Baseline	0.2474

Tables 5, 6, and 7 show the results for English and Tables 8, 9, and 10 for Japanese. The values in these tables are macro-averages over the systems for each language; individual performance for each system is shown in the appendix.

It can be seen from the tables that, for English, NCDS and RSL17BD performed well in terms of JSD and MSE. In terms of classification-related metrics, KTH, RSL17BD, and PLECO seem to have performed well. As for Japanese, except for F1(PB+B), NTTCS outperformed the other teams followed by OUARS.

6. Summary and Future work

We described our dialogue breakdown detection challenge 3. We prepared both English and Japanese datasets, and eight teams competed using methods for detecting dialogue breakdown. Although the submitted runs still to be examined in detail, we have promising results and interesting methods for dialogue breakdown detection. We aim to hold the next challenge to further improve the detection performance so that dialogue systems with fewer breakdowns can be achieved.

We want to make a final note on the quality of dialogue breakdown annotation. In Table 1, we saw considerable differences in Fleiss' κ; the values obtained by AMT were quite low. This may be because of our lack of care for unreliable workers. After the formal run, we performed a post hoc analysis of the annotations and found that some workers used the same label for most/all dialogues. After removing the data by such workers, we confirmed that Fleiss' κ could improve to 0.1-0.2 for CIC and YI. On the basis of this analysis, we will consider methods to improve the reliability of dialogue breakdown annotation for future challenges.

7. Acknowledgments

We would like to thank all of the participants for their effort in exploring new methods and submitting their runs and reports. We also thank Rafael E. Banchs, Zhou Yu, Valentin Malykh, Idris Yusupov, and Yury Kuratov for graciously providing us with datasets and chatbots to make DBDC3 possible. We also thank our sponsors, Denso IT Laboratories, Nxtremer, Honda Research Institute Japan, and NTT DOCOMO, Inc., for supporting our data collection. We also thank the Japanese Society for Artificial Intelligence (JSAI) for supporting the event.

8. References

[1] R. Higashinaka, K. Funakoshi, Y. Kobayashi, and M. Inaba, "The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics," in *Proc. LREC*, 2016, pp. 3146–3150.

[2] 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.

[3] R. Higashinaka, K. Funakoshi, M. Araki, H. Tsukahara, Y. Kobayashi, and M. Mizukami, "Towards taxonomy of errors in chat-oriented dialogue systems," in *Proc. SIGDIAL*, 2015, pp. 87–95.

[4] R. Higashinaka, M. Mizukami, K. Funakoshi, M. Araki, H. Tsukahara, and Y. Kobayashi, "Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems," in *Proc. EMNLP*, 2015, pp. 2243–2248.

[5] Z. Yu, Z. Xu, A. W. Black, and A. I. Rudnicky, "Strategy and policy learning for non-task-oriented conversational systems," in *Proc. SIGDIAL*, 2016, pp. 404–412.

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

[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] H. Tsukahara and K. Uchiumi, "System utterance generation by label propagation over association graph of words and utterance patterns for open-domain dialogue systems," in *Proc. PACLIC*, 2015, pp. 323–331.

[9] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proc. EMNLP*, 2011, pp. 583–593.

9. Appendix

Tables 17 to 13 show the results for individual systems in English. Tables 23 to 28 show the results for individual systems in Japanese.

Table 11: *Results of Classification (TickTock)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
SAM2017 run1	0.5560	RSL17BD run2	0.4242	PLECO run1	0.7848
NCDS run1	0.4920	RSL17BD run3	0.4061	PLECO run2	0.7777
NCDS run2	0.4860	PLECO run2	0.3984	RSL17BD run2	0.7703
KTH run2	0.4840	PLECO run1	0.3978	RSL17BD run3	0.7643
KTH run3	0.4480	KTH run3	0.3753	RSL17BD run1	0.7574
CRF Baseline	0.4480	SWPD run1	0.3707	KTH run1	0.7568
RSL17BD run3	0.4440	RSL17BD run1	0.3701	Majority Baseline	0.7499
RSL17BD run2	0.4360	CRF Baseline	0.3685	SWPD run1	0.7027
RSL17BD run1	0.4200	KTH run1	0.3678	NCDS run3	0.6871
SWPD run1	0.3880	Majority Baseline	0.3660	KTH run3	0.6868
PLECO run2	0.3500	NCDS run3	0.3358	CRF Baseline	0.6814
KTH run1	0.3440	SAM2017 run1	0.3076	KTH run2	0.6311
PLECO run1	0.3380	KTH run2	0.3037	NCDS run1	0.3316
NCDS run3	0.3100	NCDS run2	0.2318	NCDS run2	0.3069
Majority Baseline	0.2240	NCDS run1	0.2296	SAM2017 run1	0.2695

Table 12: *Results of JS Divergence (TickTock)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
KTH run2	0.0572	KTH run2	0.0340	RSL17BD run2	0.0328
RSL17BD run3	0.0624	RSL17BD run3	0.0430	RSL17BD run3	0.0338
RSL17BD run2	0.0634	SWPD run1	0.0442	RSL17BD run1	0.0361
RSL17BD run1	0.0664	RSL17BD run2	0.0442	KTH run2	0.0373
NCDS run2	0.0676	RSL17BD run1	0.0443	NCDS run2	0.0425
NCDS run1	0.0679	NCDS run1	0.0463	NCDS run1	0.0429
Majority Baseline	0.0736	NCDS run2	0.0464	Majority Baseline	0.0485
SWPD run1	0.0928	Majority Baseline	0.0507	SWPD run1	0.0604
PLECO run1	0.0949	PLECO run1	0.0660	PLECO run2	0.0654
PLECO run2	0.0981	PLECO run2	0.0705	PLECO run1	0.0670
NCDS run3	0.1107	NCDS run3	0.0801	SAM2017 run1	0.0695
SAM2017 run1	0.2225	SAM2017 run1	0.1716	NCDS run3	0.0797
KTH run3	0.3502	KTH run3	0.2065	KTH run1	0.1765
CRF Baseline	0.4309	KTH run1	0.2562	KTH run3	0.2568
KTH run1	0.4474	CRF Baseline	0.2803	CRF Baseline	0.3030

Table 13: *Results of Mean Squared Error (TickTock)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
KTH run2	0.0308	KTH run2	0.0393	RSL17BD run2	0.0324
RSL17BD run3	0.0357	SWPD run1	0.0514	RSL17BD run3	0.0345
RSL17BD run2	0.0361	RSL17BD run3	0.0515	KTH run2	0.0357
NCDS run2	0.0375	RSL17BD run1	0.0531	RSL17BD run1	0.0371
NCDS run1	0.0376	RSL17BD run2	0.0531	NCDS run2	0.0404
RSL17BD run1	0.0380	NCDS run1	0.0547	NCDS run1	0.0408
Majority Baseline	0.0410	NCDS run2	0.0549	Majority Baseline	0.0453
SWPD run1	0.0518	Majority Baseline	0.0609	SAM2017 run1	0.0636
PLECO run1	0.0539	PLECO run1	0.0762	SWPD run1	0.0667
PLECO run2	0.0550	PLECO run2	0.0804	PLECO run2	0.0700
NCDS run3	0.0691	NCDS run3	0.1210	PLECO run1	0.0717
SAM2017 run1	0.1093	KTH run3	0.1833	NCDS run3	0.0840
KTH run3	0.1804	SAM2017 run1	0.1922	KTH run1	0.1255
CRF Baseline	0.2178	KTH run1	0.2048	KTH run3	0.2520
KTH run1	0.2353	CRF Baseline	0.2342	CRF Baseline	0.2717

Table 14: *Results of Classification (IRIS)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
KTH run1	0.5000	KTH run1	0.6278	Majority Baseline	0.8713
PLECO run1	0.4760	PLECO run1	0.6156	KTH run1	0.8669
CRF Baseline	0.4740	Majority Baseline	0.5974	PLECO run1	0.8595
KTH run3	0.4700	CRF Baseline	0.5973	PLECO run2	0.8502
KTH run2	0.4640	PLECO run2	0.5963	RSL17BD run3	0.8345
SAM2017 run1	0.4620	NCDS run3	0.5753	NCDS run3	0.8303
PLECO run2	0.4580	KTH run3	0.5702	RSL17BD run2	0.8260
SWPD run1	0.4560	KTH run2	0.5360	RSL17BD run1	0.8238
NCDS run3	0.4560	SWPD run1	0.5045	CRF Baseline	0.7752
NCDS run2	0.4400	RSL17BD run3	0.4708	KTH run3	0.7683
NCDS run1	0.4260	RSL17BD run1	0.4659	KTH run2	0.7617
Majority Baseline	0.4260	RSL17BD run2	0.4262	SWPD run1	0.6458
RSL17BD run1	0.3700	SAM2017 run1	0.3703	NCDS run2	0.3673
RSL17BD run3	0.3620	NCDS run2	0.3290	NCDS run1	0.3539
RSL17BD run2	0.3420	NCDS run1	0.3225	SAM2017 run1	0.3021

Table 15: *Results of JS Divergence (IRIS)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
Majority Baseline	0.0470	KTH run2	0.0284	RSL17BD run2	0.0299
NCDS run1	0.0471	Majority Baseline	0.0314	RSL17BD run3	0.0304
NCDS run2	0.0473	NCDS run1	0.0314	RSL17BD run1	0.0307
KTH run2	0.0528	NCDS run2	0.0315	NCDS run1	0.0310
RSL17BD run2	0.0544	RSL17BD run2	0.0334	NCDS run2	0.0312
RSL17BD run3	0.0549	RSL17BD run3	0.0341	Majority Baseline	0.0312
RSL17BD run1	0.0553	RSL17BD run1	0.0346	KTH run2	0.0339
PLECO run1	0.0556	PLECO run1	0.0380	PLECO run1	0.0375
NCDS run3	0.0613	NCDS run3	0.0430	NCDS run3	0.0403
PLECO run2	0.0634	PLECO run2	0.0455	PLECO run2	0.0416
SWPD run1	0.0825	SWPD run1	0.0497	SWPD run1	0.0481
SAM2017 run1	0.2445	KTH run3	0.1813	SAM2017 run1	0.0743
KTH run3	0.3129	SAM2017 run1	0.1963	KTH run3	0.2205
CRF Baseline	0.4325	KTH run1	0.2140	KTH run1	0.2388
KTH run1	0.4816	CRF Baseline	0.2611	CRF Baseline	0.3389

Table 16: *Results of Mean Squared Error (IRIS)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
Majority Baseline	0.0271	KTH run2	0.0320	Majority Baseline	0.0345
NCDS run1	0.0273	NCDS run1	0.0350	RSL17BD run2	0.0347
NCDS run2	0.0275	NCDS run2	0.0351	RSL17BD run3	0.0352
KTH run2	0.0305	Majority Baseline	0.0356	RSL17BD run1	0.0354
PLECO run1	0.0325	RSL17BD run2	0.0384	NCDS run1	0.0358
RSL17BD run2	0.0326	RSL17BD run3	0.0391	NCDS run2	0.0360
RSL17BD run3	0.0328	RSL17BD run1	0.0399	KTH run2	0.0396
RSL17BD run1	0.0331	PLECO run1	0.0436	PLECO run1	0.0427
PLECO run2	0.0371	PLECO run2	0.0520	NCDS run3	0.0476
NCDS run3	0.0377	SWPD run1	0.0570	PLECO run2	0.0482
SWPD run1	0.0482	NCDS run3	0.0609	SWPD run1	0.0556
SAM2017 run1	0.1244	KTH run1	0.1557	SAM2017 run1	0.0734
KTH run3	0.1583	KTH run3	0.1624	KTH run1	0.1836
CRF Baseline	0.2139	CRF Baseline	0.2121	KTH run3	0.2094
KTH run1	0.2533	SAM2017 run1	0.2286	CRF Baseline	0.3013

Table 17: *Results of Classification (CIC)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
RSL17BD run2	0.4460	PLECO run1	0.3758	Majority Baseline	0.9847
RSL17BD run1	0.4180	Majority Baseline	0.3739	PLECO run1	0.9648
RSL17BD run3	0.3880	PLECO run2	0.3670	PLECO run2	0.9527
KTH run3	0.3220	CRF Baseline	0.3603	RSL17BD run2	0.9064
NCDS run3	0.2920	RSL17BD run2	0.3581	KTH run1	0.9057
CRF Baseline	0.2720	KTH run3	0.3520	RSL17BD run1	0.8916
SWPD run1	0.2620	KTH run1	0.3518	RSL17BD run3	0.8856
KTH run2	0.2520	SWPD run1	0.3376	SWPD run1	0.8675
KTH run1	0.2500	RSL17BD run1	0.3234	NCDS run3	0.8668
SAM2017 run1	0.2400	NCDS run3	0.3150	CRF Baseline	0.8541
PLECO run1	0.2340	SAM2017 run1	0.2871	KTH run3	0.8296
Majority Baseline	0.2300	RSL17BD run3	0.2770	KTH run2	0.8009
NCDS run2	0.2280	KTH run2	0.2755	NCDS run2	0.3584
PLECO run2	0.2220	NCDS run2	0.2418	NCDS run1	0.3327
NCDS run1	0.2200	NCDS run1	0.2296	SAM2017 run1	0.2796

Table 18: *Results of JS Divergence (CIC)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
Majority Baseline	0.0234	Majority Baseline	0.0074	NCDS run1	0.0110
RSL17BD run2	0.0246	RSL17BD run2	0.0132	NCDS run2	0.0112
NCDS run1	0.0250	RSL17BD run1	0.0145	RSL17BD run2	0.0139
NCDS run2	0.0252	NCDS run1	0.0145	Majority Baseline	0.0155
RSL17BD run1	0.0283	NCDS run2	0.0147	RSL17BD run1	0.0180
RSL17BD run3	0.0298	RSL17BD run3	0.0149	RSL17BD run3	0.0183
KTH run2	0.0481	KTH run2	0.0211	KTH run2	0.0234
NCDS run3	0.0525	NCDS run3	0.0291	NCDS run3	0.0414
SWPD run1	0.0866	SWPD run1	0.0371	SWPD run1	0.0504
PLECO run1	0.0874	PLECO run1	0.0480	SAM2017 run1	0.0632
PLECO run2	0.0927	PLECO run2	0.0512	PLECO run1	0.0697
SAM2017 run1	0.3116	KTH run3	0.1848	PLECO run2	0.0733
KTH run3	0.3574	KTH run1	0.2116	KTH run1	0.1935
KTH run1	0.4203	CRF Baseline	0.2418	KTH run3	0.2554
CRF Baseline	0.4780	SAM2017 run1	0.2423	CRF Baseline	0.3770

Table 19: *Results of Mean Squared Error (CIC)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
Majority Baseline	0.0143	Majority Baseline	0.0083	NCDS run1	0.0131
RSL17BD run2	0.0147	RSL17BD run2	0.0148	NCDS run2	0.0133
NCDS run1	0.0151	RSL17BD run1	0.0160	RSL17BD run2	0.0166
NCDS run2	0.0153	RSL17BD run3	0.0165	Majority Baseline	0.0183
RSL17BD run1	0.0170	NCDS run1	0.0167	RSL17BD run1	0.0218
RSL17BD run3	0.0179	NCDS run2	0.0171	RSL17BD run3	0.0224
KTH run2	0.0295	KTH run2	0.0255	KTH run2	0.0291
NCDS run3	0.0406	PLECO run1	0.0409	SAM2017 run1	0.0517
PLECO run1	0.0512	SWPD run1	0.0421	NCDS run3	0.0583
SWPD run1	0.0519	PLECO run2	0.0432	SWPD run1	0.0610
PLECO run2	0.0543	NCDS run3	0.0535	PLECO run1	0.0908
SAM2017 run1	0.1555	KTH run1	0.1484	PLECO run2	0.0953
KTH run3	0.1840	KTH run3	0.1549	KTH run1	0.1223
KTH run1	0.2012	CRF Baseline	0.1853	KTH run3	0.2485
CRF Baseline	0.2394	SAM2017 run1	0.2789	CRF Baseline	0.3458

Table 20: *Results of Classification (YI)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
SWPD run1	0.6120	CRF Baseline	0.0909	Majority Baseline	0.9648
Majority Baseline	0.6080	RSL17BD run1	0.0909	PLECO run2	0.9027
KTH run2	0.5660	RSL17BD run2	0.0720	PLECO run1	0.8886
CRF Baseline	0.5200	SWPD run1	0.0714	RSL17BD run1	0.8588
RSL17BD run2	0.5000	PLECO run1	0.0654	RSL17BD run3	0.8584
RSL17BD run1	0.4980	KTH run2	0.0645	RSL17BD run2	0.8574
RSL17BD run3	0.4860	PLECO run2	0.0643	KTH run1	0.8397
KTH run3	0.4480	RSL17BD run3	0.0560	SWPD run1	0.8347
NCDS run3	0.3680	NCDS run3	0.0530	NCDS run3	0.8341
SAM2017 run1	0.3660	KTH run3	0.0517	KTH run2	0.7821
NCDS run2	0.3080	NCDS run1	0.0487	KTH run3	0.7522
NCDS run1	0.3040	KTH run1	0.0476	CRF Baseline	0.7382
KTH run1	0.2560	NCDS run2	0.0360	NCDS run1	0.3652
PLECO run1	0.1320	SAM2017 run1	0.0000	NCDS run2	0.3262
PLECO run2	0.1300	Majority Baseline	0.0000	SAM2017 run1	0.0127

Table 21: *Results of JS Divergence (YI)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
Majority Baseline	0.0129	Majority Baseline	0.0054	Majority Baseline	0.0076
RSL17BD run2	0.0225	NCDS run2	0.0067	KTH run2	0.0101
RSL17BD run1	0.0229	NCDS run1	0.0068	RSL17BD run2	0.0136
RSL17BD run3	0.0234	RSL17BD run3	0.0113	RSL17BD run1	0.0140
NCDS run2	0.0249	RSL17BD run1	0.0117	RSL17BD run3	0.0148
NCDS run1	0.0249	RSL17BD run2	0.0117	NCDS run2	0.0166
KTH run2	0.0344	PLECO run1	0.0186	NCDS run1	0.0167
NCDS run3	0.0427	NCDS run3	0.0220	SWPD run1	0.0188
PLECO run1	0.0475	KTH run2	0.0232	NCDS run3	0.0339
PLECO run2	0.0554	PLECO run2	0.0255	PLECO run1	0.0398
SWPD run1	0.0609	SWPD run1	0.0441	PLECO run2	0.0457
KTH run3	0.2865	KTH run3	0.1842	SAM2017 run1	0.1152
SAM2017 run1	0.3505	KTH run1	0.2555	KTH run3	0.1336
CRF Baseline	0.4222	CRF Baseline	0.2918	CRF Baseline	0.1752
KTH run1	0.4286	SAM2017 run1	0.3407	KTH run1	0.2142

Table 22: *Results of Mean Squared Error (YI)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
Majority Baseline	0.0073	Majority Baseline	0.0067	Majority Baseline	0.0073
RSL17BD run2	0.0130	NCDS run2	0.0083	KTH run2	0.0102
RSL17BD run1	0.0133	NCDS run1	0.0084	RSL17BD run2	0.0148
RSL17BD run3	0.0136	RSL17BD run3	0.0133	RSL17BD run1	0.0155
NCDS run2	0.0146	RSL17BD run1	0.0137	SWPD run1	0.0157
NCDS run1	0.0147	RSL17BD run2	0.0139	RSL17BD run3	0.0164
KTH run2	0.0216	PLECO run1	0.0213	NCDS run2	0.0182
NCDS run3	0.0276	PLECO run2	0.0278	NCDS run1	0.0183
PLECO run1	0.0286	KTH run2	0.0290	NCDS run3	0.0389
PLECO run2	0.0329	NCDS run3	0.0353	PLECO run1	0.0476
SWPD run1	0.0366	SWPD run1	0.0501	PLECO run2	0.0555
KTH run3	0.1452	KTH run3	0.1651	SAM2017 run1	0.0597
SAM2017 run1	0.1871	KTH run1	0.1920	KTH run3	0.1078
CRF Baseline	0.2029	CRF Baseline	0.2367	CRF Baseline	0.1126
KTH run1	0.2063	SAM2017 run1	0.3612	KTH run1	0.1590

Table 23: *Results of Classification (DCM)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
NTTCS run1	0.5509	NTTCS run1	0.5476	PLECO run1	0.8054
NTTCS run2	0.5454	NTTCS run2	0.5416	PLECO run2	0.7776
NTTCS run3	0.5054	NTTCS run3	0.5141	RSL17BD run2	0.7514
RSL17BD run3	0.4945	OUARS run1	0.4785	NTTCS run3	0.7457
RSL17BD run1	0.4745	PLECO run3	0.4683	CRF Baseline	0.7289
RSL17BD run2	0.4727	OUARS run3	0.4629	PLECO run3	0.7272
CRF Baseline	0.4727	PLECO run1	0.4518	RSL17BD run1	0.7226
OUARS run1	0.4690	CRF Baseline	0.4502	OUARS run3	0.7219
PLECO run3	0.4600	PLECO run2	0.4495	OUARS run1	0.7118
OUARS run3	0.4600	OUARS run2	0.4413	RSL17BD run3	0.7114
OUARS run2	0.4454	RSL17BD run3	0.1988	NTTCS run2	0.7084
Majority Baseline	0.4145	RSL17BD run1	0.1468	NTTCS run1	0.7064
PLECO run1	0.4072	RSL17BD run2	0.1234	OUARS run2	0.7009
PLECO run2	0.4000	Majority Baseline	0.0000	Majority Baseline	0.0000

Table 24: *Results of JS Divergence (DCM)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
NTTCS run1	0.0763	NTTCS run1	0.0510	NTTCS run1	0.0457
NTTCS run2	0.0768	NTTCS run2	0.0514	NTTCS run2	0.0461
NTTCS run3	0.0833	NTTCS run3	0.0557	NTTCS run3	0.0510
OUARS run1	0.0994	OUARS run1	0.0615	OUARS run1	0.0644
OUARS run3	0.1011	OUARS run3	0.0624	OUARS run3	0.0654
OUARS run2	0.1048	OUARS run2	0.0648	PLECO run3	0.0672
PLECO run3	0.1059	PLECO run3	0.0726	OUARS run2	0.0675
PLECO run2	0.1089	PLECO run2	0.0741	PLECO run2	0.0707
PLECO run1	0.1259	PLECO run1	0.0861	Majority Baseline	0.0770
Majority Baseline	0.1380	RSL17BD run2	0.0929	PLECO run1	0.0840
RSL17BD run3	0.1782	RSL17BD run1	0.1016	RSL17BD run3	0.1071
RSL17BD run2	0.1789	RSL17BD run3	0.1016	RSL17BD run1	0.1081
RSL17BD run1	0.1804	Majority Baseline	0.1075	RSL17BD run2	0.1086
CRF Baseline	0.4100	CRF Baseline	0.2602	CRF Baseline	0.2599

Table 25: *Results of Mean Squared Error (DCM)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
NTTCS run1	0.0410	NTTCS run1	0.0527	NTTCS run1	0.0460
NTTCS run2	0.0414	NTTCS run2	0.0531	NTTCS run2	0.0465
NTTCS run3	0.0450	NTTCS run3	0.0587	NTTCS run3	0.0509
OUARS run1	0.0519	OUARS run1	0.0629	OUARS run1	0.0619
OUARS run3	0.0525	OUARS run3	0.0631	OUARS run3	0.0627
OUARS run2	0.0545	OUARS run2	0.0656	OUARS run2	0.0650
PLECO run3	0.0561	PLECO run3	0.0754	PLECO run3	0.0659
PLECO run2	0.0593	PLECO run2	0.0804	PLECO run2	0.0703
PLECO run1	0.0697	PLECO run1	0.0944	Majority Baseline	0.0705
Majority Baseline	0.0710	RSL17BD run2	0.0995	PLECO run1	0.0864
RSL17BD run3	0.0966	Majority Baseline	0.1066	RSL17BD run3	0.0998
RSL17BD run2	0.0970	RSL17BD run1	0.1079	RSL17BD run1	0.1006
RSL17BD run1	0.0974	RSL17BD run3	0.1096	RSL17BD run2	0.1018
CRF Baseline	0.2169	CRF Baseline	0.2319	CRF Baseline	0.2326

Table 26: *Results of Classification (DIT)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
NTTCS run3	0.6236	OUARS run3	0.7323	OUARS run3	0.9136
OUARS run3	0.6200	OUARS run2	0.7307	OUARS run2	0.9102
NTTCS run1	0.6181	NTTCS run3	0.7306	RSL17BD run2	0.9060
OUARS run2	0.6145	NTTCS run1	0.7287	RSL17BD run1	0.8936
NTTCS run2	0.6127	NTTCS run2	0.7258	OUARS run1	0.8896
OUARS run1	0.6072	OUARS run1	0.7165	RSL17BD run3	0.8883
PLECO run3	0.5927	PLECO run3	0.7149	Majority Baseline	0.8671
PLECO run1	0.5800	PLECO run1	0.6987	PLECO run3	0.8581
CRF Baseline	0.5563	Majority Baseline	0.6857	NTTCS run3	0.8410
PLECO run2	0.5527	PLECO run2	0.6813	CRF Baseline	0.8359
Majority Baseline	0.5218	CRF Baseline	0.6634	NTTCS run1	0.8341
RSL17BD run2	0.3854	RSL17BD run2	0.4386	NTTCS run2	0.8335
RSL17BD run3	0.3745	RSL17BD run3	0.3902	PLECO run2	0.8329
RSL17BD run1	0.3563	RSL17BD run1	0.3690	PLECO run1	0.8193

Table 27: *Results of JS Divergence (DIT)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
NTTCS run3	0.0516	NTTCS run3	0.0356	NTTCS run3	0.0319
NTTCS run1	0.0524	NTTCS run1	0.0361	NTTCS run1	0.0323
NTTCS run2	0.0528	NTTCS run2	0.0364	NTTCS run2	0.0324
OUARS run1	0.0699	RSL17BD run2	0.0428	OUARS run1	0.0412
PLECO run3	0.0730	RSL17BD run1	0.0448	OUARS run3	0.0431
OUARS run3	0.0731	RSL17BD run3	0.0478	PLECO run3	0.0448
PLECO run2	0.0778	OUARS run1	0.0522	PLECO run2	0.0476
OUARS run2	0.0801	PLECO run3	0.0539	OUARS run2	0.0478
PLECO run1	0.0892	OUARS run3	0.0557	PLECO run1	0.0573
RSL17BD run2	0.0933	PLECO run2	0.0568	RSL17BD run2	0.0605
RSL17BD run1	0.0953	OUARS run2	0.0615	Majority Baseline	0.0608
RSL17BD run3	0.1003	PLECO run1	0.0665	RSL17BD run1	0.0610
Majority Baseline	0.1063	Majority Baseline	0.0898	RSL17BD run3	0.0647
CRF Baseline	0.3741	CRF Baseline	0.2085	CRF Baseline	0.2955

Table 28: *Results of Mean Squared Error (DIT)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
NTTCS run3	0.0279	NTTCS run3	0.0337	NTTCS run3	0.0372
NTTCS run1	0.0283	NTTCS run1	0.0343	NTTCS run1	0.0378
NTTCS run2	0.0285	NTTCS run2	0.0347	NTTCS run2	0.0380
OUARS run1	0.0352	RSL17BD run2	0.0422	OUARS run1	0.0444
OUARS run3	0.0365	OUARS run1	0.0446	OUARS run3	0.0454
PLECO run3	0.0394	RSL17BD run1	0.0454	OUARS run2	0.0504
OUARS run2	0.0407	OUARS run3	0.0478	PLECO run3	0.0528
PLECO run2	0.0428	RSL17BD run3	0.0489	PLECO run2	0.0562
PLECO run1	0.0499	PLECO run3	0.0503	Majority Baseline	0.0623
Majority Baseline	0.0555	OUARS run2	0.0543	PLECO run1	0.0684
RSL17BD run2	0.0561	PLECO run2	0.0548	RSL17BD run2	0.0730
RSL17BD run1	0.0580	PLECO run1	0.0639	RSL17BD run1	0.0735
RSL17BD run3	0.0609	Majority Baseline	0.0848	RSL17BD run3	0.0775
CRF Baseline	0.1870	CRF Baseline	0.1730	CRF Baseline	0.2564

Table 29: *Results of Classification (IRS)*

Run	Accuracy	Run	F1 (B)	Run	F1 (PB+B)
NTTCS run3	0.6762	NTTCS run3	0.7475	RSL17BD run1	0.8446
NTTCS run1	0.6696	NTTCS run2	0.7379	RSL17BD run3	0.8414
NTTCS run2	0.6674	NTTCS run1	0.7379	NTTCS run1	0.8347
OUARS run3	0.6208	OUARS run3	0.7140	PLECO run2	0.8328
OUARS run1	0.6075	OUARS run1	0.6998	RSL17BD run2	0.8318
OUARS run2	0.5942	OUARS run2	0.6968	NTTCS run2	0.8307
CRF Baseline	0.5676	PLECO run2	0.6953	NTTCS run3	0.8299
PLECO run2	0.5676	PLECO run3	0.6748	OUARS run3	0.8286
PLECO run3	0.5631	PLECO run1	0.6712	PLECO run1	0.8160
PLECO run1	0.5565	Majority Baseline	0.6676	OUARS run2	0.8147
Majority Baseline	0.5011	CRF Baseline	0.6252	Majority Baseline	0.8068
RSL17BD run3	0.3458	RSL17BD run2	0.2765	PLECO run3	0.8060
RSL17BD run1	0.3392	RSL17BD run1	0.2746	OUARS run1	0.8051
RSL17BD run2	0.3192	RSL17BD run3	0.2642	CRF Baseline	0.7673

Table 30: *Results of JS Divergence (IRS)*

Run	JSD (NB,PB,B)	Run	JSD (NB,PB+B)	Run	JSD (NB+PB,B)
NTTCS run2	0.0784	NTTCS run2	0.0526	NTTCS run2	0.0487
NTTCS run1	0.0787	NTTCS run1	0.0527	NTTCS run1	0.0490
NTTCS run3	0.0808	NTTCS run3	0.0553	NTTCS run3	0.0497
OUARS run1	0.0979	OUARS run1	0.0715	OUARS run1	0.0598
OUARS run3	0.0994	OUARS run3	0.0738	OUARS run3	0.0604
OUARS run2	0.1054	RSL17BD run3	0.0759	OUARS run2	0.0634
PLECO run2	0.1088	PLECO run3	0.0773	PLECO run2	0.0665
PLECO run3	0.1088	RSL17BD run2	0.0785	PLECO run3	0.0682
PLECO run1	0.1213	PLECO run2	0.0786	PLECO run1	0.0768
Majority Baseline	0.1453	OUARS run2	0.0792	Majority Baseline	0.0873
RSL17BD run3	0.1844	RSL17BD run1	0.0820	RSL17BD run1	0.1203
RSL17BD run1	0.1861	PLECO run1	0.0894	RSL17BD run2	0.1210
RSL17BD run2	0.1861	Majority Baseline	0.1201	RSL17BD run3	0.1222
CRF Baseline	0.3713	CRF Baseline	0.2451	CRF Baseline	0.2838

Table 31: *Results of Mean Squared Error (IRS)*

Run	MSE (NB,PB,B)	Run	MSE (NB,PB+B)	Run	MSE (NB+PB,B)
NTTCS run2	0.0420	NTTCS run2	0.0491	NTTCS run2	0.0566
NTTCS run1	0.0422	NTTCS run1	0.0493	NTTCS run1	0.0569
NTTCS run3	0.0435	NTTCS run3	0.0521	NTTCS run3	0.0580
OUARS run3	0.0517	OUARS run3	0.0669	OUARS run3	0.0669
OUARS run1	0.0525	OUARS run1	0.0671	OUARS run1	0.0681
OUARS run2	0.0543	OUARS run2	0.0714	OUARS run2	0.0692
PLECO run3	0.0595	PLECO run3	0.0743	PLECO run2	0.0769
PLECO run2	0.0609	RSL17BD run3	0.0754	PLECO run3	0.0779
PLECO run1	0.0684	PLECO run2	0.0793	PLECO run1	0.0893
Majority Baseline	0.0763	RSL17BD run2	0.0794	Majority Baseline	0.0923
RSL17BD run3	0.1083	RSL17BD run1	0.0821	RSL17BD run1	0.1358
RSL17BD run1	0.1094	PLECO run1	0.0902	RSL17BD run3	0.1369
RSL17BD run2	0.1107	Majority Baseline	0.1121	RSL17BD run2	0.1372
CRF Baseline	0.1950	CRF Baseline	0.2194	CRF Baseline	0.2532