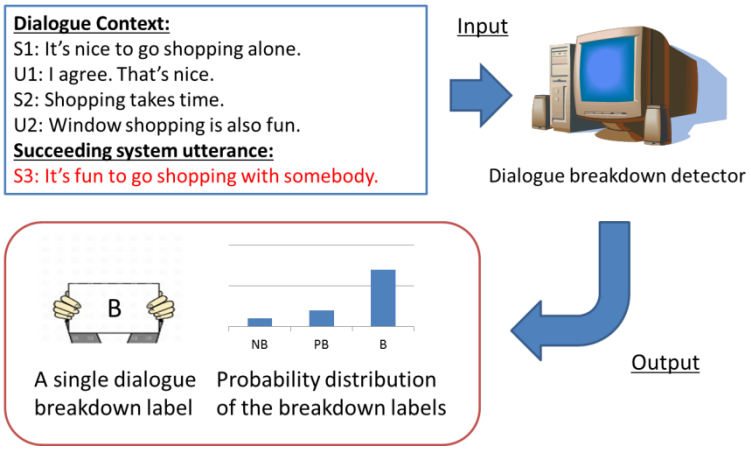
Proposal for a dialogue breakdown detection track

1) The names and affiliations of the organizers

1. Ryuichiro Higashinaka (NTT)
2. Kotaro Funakoshi (HRI-US)
3. Michimasa Inaba (Hiroshima City University)
4. Yuiko Tsunomori (NTT Docomo)
5. Tetsuro Takahashi (Fujitsu)

2) A description of the task, with particular reference to its relevance for the dialog community

The task of dialogue breakdown detection is to detect whether the system utterance causes dialogue breakdown (a situation in a dialogue where users cannot proceed with the conversation) in a given dialogue context. The participants of the dialogue breakdown detection track will 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 (See below):



The task of dialogue breakdown detection is relevant for the dialogue community because, although voice agent services are beginning to appear on the market, the limited capabilities of these systems mean that humans and machines still cannot converse as naturally as two humans. The main problem is that systems typically make inappropriate utterances that lead to dialogue breakdowns. The dialogue breakdown detection technology will be useful for chat-oriented dialogue systems in which continuing the conversation is important. The technology will also be useful for the recovery from errors. More details of the task can also be found in [1].

3) A description of the data that will be provided for participants

In the past two years, we have hosted two challenges regarding dialogue breakdown detection in Japan [2,3]. For these events, we have created the following data which are already publicly available.

1. 100 chat-oriented dialogues (each dialogue contains 21 utterances) with dialogue breakdown annotation by 24 annotators for each system utterance.
2. Approx. 1000 chat-oriented dialogues with dialogue breakdown annotation by 2-3 annotators for each system utterance.
3. 300 chat-oriented dialogues (collected with three different systems) with dialogue breakdown annotations by 30 annotators for each system utterance.

Unfortunately, the data above are in Japanese; however, we are currently in the process of collecting 100 dialogues in English and annotating them with dialogue breakdown labels. The data will also be available in a few months.

In the previous challenges, we had 7 and 8 participating teams, respectively. We would like to take this opportunity to internationalize the event for realizing dialogue breakdown detection in multiple languages.

4) A description of the evaluation methods that will be used to compare peer systems

1. Classification-related metrics: Accuracy, Precision, Recall and F-measure
   These metrics are calculated by comparing the output of the detector and the gold label determined by majority voting.
2. Distribution-related metrics: JS Divergence and Mean squared error
   These metrics are calculated by comparing the predicted distribution of the labels with that of the gold labels.

5) References

1. Higashinaka et al. "The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics." of the Language Resources and Evaluation Conference. 2016.
2. Website for the 1st dialogue breakdown detection challenge (In Japanese)
   https://sites.google.com/site/dialoguebreakdowndetection
3. Website for the 2nd dialogue breakdown detection challenge (In Japanese)
   https://sites.google.com/site/dialoguebreakdowndetection2/

Examples of the data and annotations
(B: Breakdown, PB: Possible Breakdown, or NB: Not a breakdown)

A dialogue example with few breakdowns:

	Utterance	NB	PB	B
S	Hello. I was waiting for you to talk to me.	24	0	0
U	*Really?! I'm so happy to hear that.*			
S	What are you doing for your health?	24	0	0
U	*Well, I make it a rule to drink green juice every day.*			
S	What kind of food do you like to eat?	23	1	0
U	*I love pizza!! What's more, I love Korean BBQ too.*			
S	Do you like eating?	24	0	0
U	*Love it so much. I'd like to eat as much as I like if I would never get fat.*			
S	Speaking of Korean BBQ, do you like salted ox tongues?	23	1	0
U	*I like boneless short ribs better than salted ox tongues. Boneless short ribs are so good, don't you think?*			
S	You eat them at BBQ restaurants, do you?	22	1	1

A dialogue example with many breakdowns:

	Utterance	NB	PB	B
U	*You'd better stay at home, as it is very hot outside.*			
S	You have sweat on your face, don't you?	22	2	0
U	*Indeed.*			
S	So you do.	24	0	0
U	*I hate to sweat.*			
S	Your make-up gets smeared due to sweat, doesn't it?	24	0	0
U	*I don't wear make-up.*			
S	I want make-up.	9	10	5
U	*I don't want it.*			
S	Thanks for the meal!	1	5	18
U	*Did you eat anything?*			
S	The make-up is appropriate.	3	6	15