Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 54 (2019) 31–43

www.elsevier.com/locate/csl

Dialogue breakdown detection robust to variations in annotators and dialogue systems☆

Junya Takayama*, Eriko Nomoto, Yuki Arase

Graduate School of Information Science and Technology, Osaka University, Yamadaoka 1-12, Suita City, Osaka Prefecture 565-0871, Japan

Received 11 March 2018; received in revised form 19 July 2018; accepted 26 August 2018
Available online 29 August 2018

Abstract

Dialogue breakdown is a significant problem in conversational agents. Timely breakdown detection helps the agents quickly recover from mistakes, minimizing the impact on user experience. In this paper, we focus on two problems: variations in determining a response that breakdowns a conversation *i.e.*, subjectivity, and variations in breakdown types due to designs of conversational agents, *i.e.*, variationality. To address the subjectivity, which decreases the agreement rate among annotators, our methods detect a dialogue breakdown by ensembling detectors trained by different sets of annotators that are grouped using a clustering algorithm. To address the variationality, our methods apply two types of detector architectures to capture global and local breakdowns. The long short-term memory detector considers the global context and the convolutional neural networks detector is sensitive to the local characteristics. The ensemble of all detectors makes a final judgment. The results of the Japanese task in the Dialogue Breakdown Detection Challenge 3 (DBDC3) confirm that our approach significantly outperforms the baseline, which uses the conventional conditional random field. Detailed error analysis reveals that our encoders based on a convolutional neural network and a long short-term memory have different characteristics. It also confirms the effects of annotator clustering.

Keywords: Dialogue breakdown detection; Ensemble learning; Clustering; Convolutional neural network; Recurrent neural network

1. Introduction

Conversational agents with chit-chat abilities, which are known as chat-bots, are becoming popular. Although they are generally implemented by generation-based or example-based approaches, the fact that output utterances often collapse the dialogue context remains a challenge. In an effort to overcome this obstacle, the Dialogue Breakdown Detection Challenge (DBDC) (Higashinaka et al., 2016) holds a shared task to detect inappropriate utterances, which cause breakdowns in user-system dialogue.

☆ This is an extended version of our paper (Takayama et al., 2017).
This paper has been recommended for acceptance by Roger K. Moore

* Corresponding author.
E-mail address: takayama.junya@ist.osaka-u.ac.jp (J. Takayama).

https://doi.org/10.1016/j.csl.2018.08.007
0885-2308/

There are two challenges in dialogue breakdown detection: *subjectivity* and *variationality*. The former results in variations when determining if a response causes breakdowns in a conversation, which lowers the agreement rate among annotators. The latter is due to the design of conversational agents, which lead to variations in the breakdown types. For the subjectivity, judgments on dialogue breakdowns are inevitably subjective. DBDC data consists of user-system dialogue, where each system's utterance is annotated with breakdown labels (O: *Not a breakdown*, T: *Possible breakdown*, X: *Breakdown*) by some annotators. The distribution of annotated labels could be biased among annotators, *i.e.*, one annotator group is more sensitive while another one is more generous. In fact, for development and test data of DBDC, Fleiss's Kappa, which is an index to measure the level of annotation agreements, is as low as 0.14 to 0.36, respectively. Table 1 shows examples extracted from the DBDC data. The example in the last row indicates that the system responded with an irrelevant topic to the user's utterance. The annotators' agreement is high because such a case is easy to judge. However, not all cases are clear. The first two rows show cases where the annotators' judgments are split. Annotators who regarded the first example as a breakdown may have thought that the system should have responded with something about a zoo rather than an aquarium. As for the second example, annotators may have considered that the user did not like to talk about business manners. These examples show that the annotators' judgments can be controversial when assessing dialogue breakdown.

For the variationality, different chat-bots show different characteristics in their responses. DBDC data consists of dialogues collected from three systems: a chat-bot API provided by NTT Docomo (DCM), Denso IT Laboratories system (DIT), and an IR-status based system (Ritter et al., 2011) (IRS). Differences in their responses are easily characterized by the sentence length. The average sentence lengths of DCM, DIT and IRS are 7, 29, and 13, respectively. We observed that dialogue breakdowns are caused by mismatches when considering the global and local context in a user's utterance and a system's response.

Given these observations, we propose two simple yet effective approaches to detect a dialogue breakdown for each problem.

- For subjectivity, we propose methods to consider subjectivities in annotations that result in different distributions of annotation labels. First, the *k*-means clustering is employed to cluster annotators based on their annotation distributions. Second, a dialogue breakdown detector is constructed for each cluster using the annotation labels. Finally, a breakdown probability is estimated by ensembling predictions of all detectors, *i.e.*, taking the average of all predictions.
- For variationality of the conversational agents, our approach uses three models as detectors that employ deep neural networks (DNNs). One model uses two series Long Short-Term Memory (LSTM) encoders, one uses two parallel Convolutional Neural Network (CNN) encoders, and the other ensembles these detectors.

We participated in the DBDC3 (Higashinaka et al., 2017) Japanese task. The F-score of our approach is 63.6% for breakdown detection. It outperforms the baseline detector based on conditional random fields (CRFs) by 5.6%. This confirms the effectiveness of our approach, which considers both the subjectivity in annotations and the variationality of conversational agents. This paper provides a detailed analysis of the breakdown detection results, different characteristics of the detectors using CNN and LSTM, the effect of annotation clustering, and their common challenges to further improve the breakdown sensitivity.

Table 1
Examples of utterance-reply pair and dialogue breakdown annotations.

user utterance	system reply	annotation O	T	X
よく動物園に行ったりするんですか？ (Do you often visit a zoo?)	水族館だね。 (I'm going aquariums.)	7	17	6
あだ名、教えてください (Tell me your nickname.)	職場で同僚をあだ名で呼ぶのは、どうかと思うよ。 (I don't think it's appropriate to call colleagues by their nicknames at the office.)	8	13	9
私も眠くなってきました (I'm sleepy, too.)	チキンライス！あれ？そんなメニューなかったなぁ。 (A chicken pilaff! I don't think there was such a menu.)	0	1	29