Utterance Selection based on Sentence Similarities and Dialogue Breakdown Detection on NTCIR-12 STC Task

Hiroaki Sugiyama
2-4, Hikari-dai, Seika-cho, Souraku-gun, Kyoto, Japan
NTT Communication Science Laboratories
sugiyama.hiroaki@lab.ntt.co.jp

ABSTRACT

This paper describes our contribution for the NTCIR-12 STC Japanese task. The purpose of the task is to retrieve tweets that suits as responses of a chat-oriented dialogue system from a huge number of tweets pool. Our system retrieves tweets based on following two steps: first it retrieves tweets that resemble to input sentences, and then, it filters inappropriate tweets in terms of the dialogue flow naturalness using a dialogue breakdown detection system. Our experiments show that although the dialogue breakdown detection cannot distinguish best and medium appropriateness, it works well even in data domains that are slightly different from expected ones.

Team Name

NTTCS

Subtasks

Short Text Conversation (Japanese)

Keywords

dialogue breakdown detection, utterance selection, chat-oriented dialogue

1. INTRODUCTION

The NTCIR-12 STC Japanese task aims to retrieve sentences from a huge sentence pool that are suitable as responses for a dialogue system. Given a tweet as a user utterance that is randomly selected from tweet corpus, each participating system returns 10 tweets that are expected as appropriate responses of the input user utterance.

Our system on NTCIR-12 STC Japanese task retrieves tweets based on the combination of retrieval-based and filtering-based approach. Retrieval-based approach retrieves tweets that are similar to input user utterances from a huge tweet pool. Although this approach can return tweets that relate to the rough topics of user utterances such as foods or travels, most system utterances are not appropriate in terms of the detailed topics, since most tweets are too complicated to calculate the "similarities" between utterances. To filter tweets that are inappropriate as responses in terms of dialogue naturalness, we adopt our previously proposed dialogue breakdown detection (*DBD*) method [4]. The DBD method detects inappropriate utterances that cause dialogue breakdown in casual conversation between human and a dialogue system.

2. RELATED WORK

For utterance selection for open-domain dialogue, Ritter et al. [3] proposed IR-status, IR-response. Unlike web documents, Twitter contains many short conversational sentences. On Twitter, since users often post sentences related to their daily lives and chat using its in-reply-to function, these sentences are written about daily topics in light, breezy styles, making them very suitable for conversational system's utterances. Focusing on these features, Ritter et al. proposed three approaches: IR-status, IR-response [3]. IR-status retrieves reply posts whose associated source posts most resemble user utterances. This approach is reasonable to leverage the in-reply-to function; however, when it cannot find similar sentences or the relation between source and reply posts depends on unobserved contexts, it generates irrelevant, incomprehensible sentences as system utterances. The IR-response approach resembles IR-status, but it retrieves reply posts that most closely resemble user utterances. Even though this approach avoids generating irrelevant utterances, IR-response has difficulty expanding the conversation topics; if the same sentence as the user utterance is contained in the corpus, it parrots the user utterance.

3. OUR SYSTEM

If we retrieve tweets based on only sentence similarities, the retrieved tweets consist of a few appropriate tweets and many inappropriate tweets. To filter such inappropriate tweets, we adopt dialogue breakdown detection (*DBD*) [4], which detects inappropriate utterances that cause dialogue breakdown in casual conversation. Thus, our approach is based on two steps as follows: first, we retrieve several tweets more than the final output numbers that are expected to be roughly related to the input tweets. Thus, we filter the retrieved tweets using the DBD system to collect final output tweets.

3.1 Sentence similarity based retrieval

To calculate similarities between sentences, first we define a sentence vector v_s for each sentence s based on a word embedding method like word2vec [2]. Here, we utilize a mean vector of the words of a sentence as follows:

$$v_s = \frac{\sum_{w \in W_s} v_w}{|W_s|}, \quad (1)$$

where v_w is a vector of word w and W_s is the word set of s, and $|W_s|$ is the number of words in s. This is a basic approach to define a sentence vector using word embedding methods.

To calculate similarities between sentences, We also define a similarity between sentences $Sim(s_1, s_2)$ as inner products

as follows

$$Sim(s_1, s_2) = v_{s_1} \cdot v_{s_2} \quad (2)$$
$$= \sum_i v_{s_1}(i) v_{s_2}(i), \quad (3)$$

where $v_s(i)$ shows the value at index i of the sentence vector s. After the similarity calculation, we choose N_{ret} tweets that have high similarities for each input tweet.

In this study, we utilize word2vec as the word embedding method, and set N_{ret} with 20.

3.2 Dialogue breakdown detection based filtering

To filter inappropriate tweets that are gathered by the sentence similarity based approach, we adopt our previously proposed dialogue breakdown detection (*DBD*) system, which detects inappropriate utterances that cause dialogue breakdown in casual dialogue. Our DBD system is the state-of-the-art one in the DBD challenge conducted by Higashinaka et al. [1], where some DBD systems tried to detect inappropriate utterances generated by a dialogue system, it promises to work well to detect inappropriate tweets also in this NTCIR-12 STC task.

Our DBD system is developed using a deep multilayer perceptron classifier, which is trained with a dialogue corpus distributed by Higashinaka et al [1]. The corpus consists of about 1046 text-chat dialogues between humans and systems[1]. Each dialogue has 10 human utterances and 11 system utterances by turns. All the system utterances are evaluated by two human annotators whether the utterance seems to cause dialogue breakdown as follows: O (correct), T (cannot be judged), X (breakdown).

The settings of the deep multilayer perceptron classifier are as follows: the number of neurons of each hidden layer are 7500, 7500, 5000, 2500 (input to output), the activation functions are ReLU, the output function is softmax, optimization methods is AdaGrad with 50% dropout. We adopt f-value of T+X estimation, where we think T and X are the same annotations, as the criterion of the model selection. Table 1 shows the features of the deep multilayer perceptron classifier. The word class features are calculated using k-means of vectors obtained from word2vec. Since the DBD system can return the probability of the dialogue breakdown, we utilize the probability for the reranking of retrieved candidate tweets.

Table 1: Feature descriptions

Feature	Description
Word	Bag-of-words of input and candidates
Word class	Word class of input and candidates
Word comb.	The combination of co-occurred words
Similarity	Cosine similarities between input and candidates
Perplexity	Perplexity of input and candidates

4. EXPERIMENTS

4.1 Experiment settings

In the NTCIR-12 STC Japanese task, we submitted two runs: *without DBD reranking* and *with DBD reranking*. *Without DBD reranking* is our baseline method, in which we submitted tweets that are retrieved based only on the sentence similarities. *With DBD reranking* is our proposed method that we use DBD based reranking in addition to the sentence similarity based retrieval. Through this comparison, we examine the effectiveness of the DBD system.

We submit best 10 tweets for each input tweet in a single run. The evaluation of the tweets are performed by 10 people who annotate one of 0, 1, 2 for each tweet (2 is best).

4.2 Results

Table 2: Annotated scores by human annotators for submitted two runs

Appraoch	2-1	2-5	12-1	12-5
W/o DBD reranking	0.0921	0.0698	0.2639	0.2318
W/ DBD reranking	0.0876	0.0677	0.2946	0.2333

Table 2 shows the accuracy of the two runs. Here, X-Y represents the evaluation settings; X=2 means label "2" is regarded as correct, while X=12 means labels "1" and "2" are regarded as correct. Y=1 means only rank-1 replies are evaluated, while Y=5 means replies with rank <= 5 are evaluated. Table 2 shows that the two approach with and without DBD reranking are almost similar. In accuracy 2-1 and 2-5, with DBD reranking is slightly worse than the without DBD reranking; contrary to this, with DBD reranking is better in accuracy-12-1 result. This indicates that our DBD system does not have enough sensitivity to distinguish labels "1" and "2". There are two reasons of the low sensitivity. One is that we train the DBD system using dialogue system's utterances that are less appropriate than the tweets written by human. The other is that we train the DBD with the same setting as the X=12. Considering the differences of the data types, the result is reasonable one. This indicates that the DBD system works well even when the data domains are not identical.

5. CONCLUSIONS

We propose a sentence retrieval system for chat-oriented dialogue systems to respond to open-domain user utterances. Our system retrieves tweets from a huge tweet pool with following the two steps: sentence similarity based retrieval and dialogue breakdown detection *DBD* based filtering. Our experiments show that the DBD works well even in unexpected data domains.

6. REFERENCES

[1] R. Higashinaka, K. Funakoshi, Y. Kobayashi., and M. Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *10th edition of the Language Resources and Evaluation Conference*, 2016 (to appear).

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 1–9, 2013.

[3] A. Ritter, C. Cherry, and W. Dolan. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, 2011.

[4] H. Sugiyama. Chat-oriented Dialogue Breakdown Detection based on Abstract :. In *SIG-SLUD*, pages 51–56, 2015 (in Japanese).

[1] `https://dev.smt.docomo.ne.jp/?p=docs.api.page&api_name=dialogue&p_name=api_usage_scenario`