# YUILA at the NTCIR-12 Short Text Challenge: Combining Twitter Data with Dialogue System Logs

Hiroshi Ueno
Yamagata University, Japan
tmk56575@st.yamagata-u.ac.jp

Takuya Yabuki
Yamagata University, Japan
tem99814@st.yamagata-u.ac.jp

Masashi Inoue
Yamagata University, Japan
mi@yz.yamagata-u.ac.jp

## ABSTRACT

The YUILA team participated in the Japanese subtask of the NTCIR-12 Short Text Challenge task. This report describes our approach to solving the responsiveness problem in STC task by using external dialogue log corpus and discusses the official results.

## Team Name

YUILA

## Subtasks

Short Text Conversation (Japanese)

## Keywords

Twitter, microblog, utterance pair, dialogue log

## 1. INTRODUCTION

The YUILA team participated in the Japanese task of the NTCIR-12 Short Text Conversation (STC) task.This report describes our approach to solving the STC problem and discusses the official results.

The STC task requires the system to find the most relevant short text within a set of short text given an input text which is also a short text. The task resembles to the realization of example-based dialogue systems. In example-based dialogue systems, the simplest method to find the relevant response given an input from the user is by selecting the most similar example in terms of a specified feature and a similarity metric. The problem in such an approach is that even though the selected example is semantically similar to the input text, it is not guaranteed that the selected text is relevant as a response. Being a response to an utterance requires to be semantically coherent to the input or the post utterance. Also responses must entail the functionality as the reaction to the post utterances. To avoid the irrelevance as the response, there are two approaches. The first is the use of knowledge on the degree of responsiveness; the knowledge can be used either as the filtering rules or the weight for re-ranking. The second is the use of the existing post and response relationship between texts; if some utterances are known to be the responses to other utterances, they are considered as relevant in terms of responsiveness. We thought the first approach might be difficult by looking only at surface text. Therefore, we took the second approach. The second approach, the utilization of the post-response relationship, has been tested for Twitter data. However, in reality, there are few pairs that can automatically be detected as post-response pairs, probably due to the monologue nature of the Twitter service. Instead, we use the external dialogue resource that has sufficient post-response pairs. We used the dialogue break down corpus that has been created by recording the utterance logs between users and a dialogue system. Although the size of the corpus is small and the possibility that there will be a semantically coherent utterance found is also small, we hoped the introduction of the external resource in this way could improve the task scores.

## 2. METHOD

We employed two approaches. The first is seeking the response to the input text within the short text or tweet corpus that is specified by the task organizer. The second is utilizing the external dialogue corpus that has clear post and response relationship. We used a corpus that consists of the dialogue logs made between human and machine in Japanese[2] [1]. In both approaches, we employed tfidf weighting to create feature vectors for each tweet and cosine measure to calculate similarity scores. We used the following formula for the tfidf weighting of term $t$ when tf represent term frequency, df indicates document frequency, idf means inverse document frequency, and the set of documents is $D$:

$$\text{tfidf} = \text{tf} * (1 + \text{idf})$$
$$\text{where} \quad \text{idf} = \log \frac{|D| + 1}{\text{df} + 1}$$

The calculated tfidf scores were then normalized. Ideally, the document set $D$ should be the entire tweets but for computational efficiency, we divided the whole corpus into subsets of $10,000$ text and calculated idf scores for each of them. The tfidf weighted vectors for documents 1 and 2 are represented as $\mathbf{v}_1$ and $\mathbf{v}_2$, the similarity between the two documents were calculated as follows:

$$\text{cossim} = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{|\mathbf{v}_1||\mathbf{v}_2|}$$

The motivation of the latter approach can be explained by the following example. When an input text is "I want to play the game more." and the most similar tweet in the data set is "This is the game I want to play more.", the latter text is mentioning the similar topic to the input text but not considered as relevant as the response. Instead, when the similar utterance is sought in the external dialogue log corpus, in which the utterances always followed by the responses, we

[1]https://sites.google.com/site/dialoguebreakdowndetection/chat-dialogue-corpus

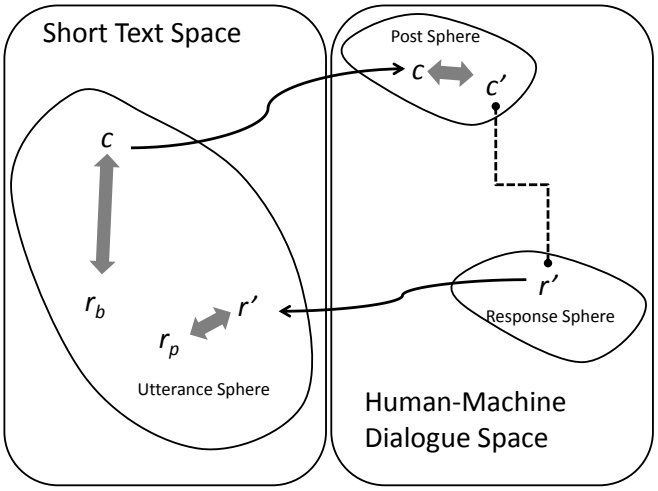


**Figure 1: Schematic explanation of proposed method**

can expect that the post and response relationship can be automatically obtainable. For example, when the most similar utterance to the above input text is "The game is over. I don't want to do this anymore" in the external corpus, the next utterance "I understand." made by the other speaker is used as the pseudo candidate response. Then, the most similar tweet to the pseudo candidate response is sought. For example, the tweet "I see." is used as the response. Comparing the response tweet selected within the twitter corpus, "This is the game I want to play more.", the tweet selected after going through the post and response relationships in the external corpus, "I see." seems more relevant response to the input "I want to play the game more."

This concept is schematically depicted in Figure 1. In the baseline method, the input text $c$ represented as a word vector weighted by tfidf scores are compared with all candidate utterances and the tweet with the highest similarity score is selected as the response $r_b$. In the proposed method, the input text $c$ is compared with utterances in the external corpus. The closest utterance $c'$ is selected as the alternative input. Then, the response to $c'$, $r'$ is determined. The response $r'$ in the external corpus is compared with the tweets and the most similar tweet $r_p$ is used as the final response.

## 2.1 Preprocessing of Twitter Data

Among $1,000,000$ tweets that were specified by the task organizers, we could retrieve $925,659$ after crawling. Other tweets were not accessible. The stored tweets were phonologically analyzed. By skimming through the data, we found that short tweets were often irrelevant as responses. By applying the filtering rules, 19% of the retrieved tweets $(171,484)$ were removed and $750,813$ were used.

The filtering rules were as follows:

1. Remove user names at the beginning of the tweets (username).
2. Remove tweets that were less than 8 words excluding user names at the beginning of the text.

The username strings removed in the first step were reverted after preprocessing. The reason usernames were removed in the first step is that they increased the number of words in text in an undesirable manner. The twitter usernames were regarded as unknown words and were divided into irregular fragments during the morphological analysis. Removing username strings have two effects. The positive effect is that the two tweets with and without usernames are considered equally similar to a query tweet; the semantic content of tweet were considered. The negative effect is that the symbol at the beginning of the tweets are considered as the sign of response rather than posts. Such tweets should be prioritized in the STC task.

## 2.2 Preprocessing of Dialogue Data

From the external dialogue corpus, $11,460$ utterance pairs were extracted. The utterances made by the users were compared with input tweet and the ones by the dialogue system were used as the seed $r'$ in Figure 1. We considered that the utterances made by the system are more likely to be considered as the response than the ones by users. When using the external corpus that contains numerous dialogue breakdowns, we should remove the broken utterance pairs from our database. For the purpose, we eliminated utterances that more than half annotators judged as broken. By applying the filtering rules, 9% of the pairs (857) were removed and $10,603$ were used.

Even though these apparently irrelevant utterance-response pairs could be removed, there are other utterance pairs that are not suitable for our method. One example is the utterances for topic shift. The dialogue system can change the dialogue topic by using the phrase "by the way". However, such utterances are not relevant in the STC task that considers single response made by the system. Other example is the back-channeling utterances for acknowledgment. Dialogues systems can encourage users to continue their utterances but in STC task, the system may be expected to respond with something meaningful.

# 3. RUNS

We have submitted four runs. The run1 is our baseline method in which the response is sought within the twitter data set. The run2 is our proposed method in which the external dialogue corpus is utilized. The run3 and the run4 are the extension of the run2 by combining the output of run1 and run2 in different ways. The run1 and run2 contains five tweet items with highest scores. The run3 and run4 combines the two sets of five items and duplicated items were removed.

**run1** Similarity search within the twitter sphere (5 items).

**run2** Similarity search with in the Post-Response space (5 items).

**run3** Combination of run1 and run2 outputs prioritizing run1 results ($< 10$ items).

**run4** Combination of run1 and run2 outputs prioritizing the overlapped items in both runs ($< 10$ items).

Participants are allowed to submit ten tweets per query for each run. However, we could prepare only five tweets for run1 and run2. The tweets that had become inaccessible on 4th February 2016 are not allowed to be included in the submission. We removed those tweets from the submissions

**Table 1: Summary of the results**

| Evaluation | Run | Accuracy |
|---|---|---|
| 2-1 | 1 | 0.1470 |
| | 2 | 0.0649 |
| | 3 | 0.0649 |
| | 4 | 0.0663 |
| 2-5 | 1 | 0.1267 |
| | 2 | 0.0567 |
| | 3 | 0.0568 |
| | 4 | 0.0568 |
| 12-1 | 1 | 0.3480 |
| | 2 | 0.2485 |
| | 3 | 0.2485 |
| | 4 | 0.2490 |
| 12-5 | 1 | 0.3087 |
| | 2 | 0.2254 |
| | 3 | 0.2254 |
| | 4 | 0.2254 |

after assembling lists. Therefore, some submissions contain less than five or ten items. Four items in run1 and seven items in run2 were removed in our submissions.

# 4. RESULTS

The results of our runs are summarized in Table 1. The evaluation columns are described in X-Y format where X is the criteria of relevance and Y is the number of items evaluated from the top of the output ranking. When X is set to be 2, label 2 is regarded as correct, while X is set to be 12, both label 1 and 2 are regarded as correct. When Y is set to be 1, only rank-1 replies are evaluated, while Y is set to be five, replies with rank smaller or equal to 5 are evaluated.

The top five items in run3 are the same as the run2 and when only five top five items were used for the evaluation, the accuracy for the run2 and the run3 must be the same. The reason there is a slight difference between accuracy scores for run2 and run3 for the 2-5 condition is that run2 contains only four items after removal of the unusable tweet but run3 contains five with the fifth item coming from the run1.

Our proposed method, run2, obtained far lower accuracy scores in all conditions and the gain obtained by the output combination, run3 and run4, are marginal.

As the gap in accuracy scores suggest, there are few differences in the outputs of run3 and run4. There are only three differences over all queries as summarized in Table 2. For the first query in the table, response “Please! Please!” is scored higher than “Please please please” by the human annotator and run4 was obtained higher accuracy for the 2-1 evaluation condition. For the second query, the output “Done. Please” with fewer score 2 annotations than “Done. Please.” took the first position for run4 and got lower score for the 2-1 evaluation condition but slightly higher score for the 12-1 condition. For the third query, the position of one item was different but did not affect the accuracy scores. The different methods of output combination yielded these marginal changes in the final ranked lists.

# 5. DISCUSSION

The use of dialogue log data performed far worse than the simple similarity search within the candidate tweets. The failure of run2 indicates that the importance of responsiveness to be a response to a short text is relatively small but the semantic coherence to an input text is. Therefore, for the performance improvement, the investigation of feature or representation of short text and the similarity metrics are considered important.

The external corpus was used to obtain post and response pairs beyond the initial corpus to obtain varieties of expressions. Therefore, it is possible to use externally collected tweet-reply pairs as the source of dialogue examples. However, Twitter is not rich with conversational content. Higashinaka et al. found that only 2.6% of tweets are conversational[1]. We think that one of the most important differences between the Chinese and Japanese subtasks except the language difference is the difference of communication style. As shown in the Figure 2 of the overview paper of the STC task[7], Weibo users seem to communicate by commenting to a post. On the other hand, majority of Twitter users do not respond to posts often and the users post something without expecting reactions. Therefore, for the Japanese subtask, we consider the use of external corpus other than Twitter seems promising. Outside the Twitter sphere, we can use the BBS data as corpus for particular writing style [5]. In BBS services we can find some anchor text that refer to past posts made in the service and use them as the dialogue data. Similarly, we can use community question answering (CQA) corpus for finding particular type of responses [3]. If we want to have more serious conversation than the average online chatting, we can consider using task-oriented dialogue corpus such as counseling [4]. When task-specific or domain-specific corpus was introduced, there may need be a conversion process from the original corpus to reusable portable format [6]. The comparison of the influence given by the different external corpus is an interesting future direction.

An important limitation of our approach is that the method used did not take the correspondence between posts and responses into account except the word-level semantics. We can consider speech act, sentiment, entity relation, and discourse structure as stated in [7]. The classification of queries into types may be the first step toward that direction.

# 6. CONCLUSIONS

We have conducted the STC task based on the sentence similarity search. The search space was either within the short text data set (Tweets) or both short text data set and the external dialogue corpus. The use of external resource deteriorates the accuracy of selected utterances as responses.

# 7. REFERENCES

[1] R. Higashinaka, N. Kawamae, K. Sadamitsu, Y. Minami, T. Meguro, K. Dohsaka, and H. Inagaki. Building a conversational model from two-tweets. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 330–335. IEEE, 2011.

[2] R. Higashinaka, M. Mizukami, K. Funakoshi, M. Araki, H. Tsukahara, and Y. Kobayashi. Fatal or not? finding errors that lead to dialogue breakdowns in

**Table 2: Comparison of run3 and run4 results**

| Query<br>Query (in Japanese) | Run | Position | # of Score 2 | # of Score 1 | Output<br>Output (in Japanese) |
|---|---|---|---|---|---|
| @souROSE523689 Please<br>@souROSE523689 \n よろしくお願いします！ | run3 | 1 | 2 | 8 | Please please please<br>お願いしますお願いしますお願いします |
| | | 2 | 6 | 4 | Please! Please!<br>お願いします！お願いします！ |
| | run4 | 1 | 6 | 4 | Please! Please!<br>お願いします！お願いします！ |
| | | 2 | 2 | 8 | Please please please<br>お願いしますお願いしますお願いします |
| @Aya_bianNights Please<br>@Aya_bianNights よろしくお願いします。 | run3 | 4 | 2 | 6 | Done. Please<br>しました、よろしくお願いします |
| | | 5 | 1 | 8 | Done. Please.<br>しました。よろしくお願いします。 |
| | run4 | 1 | 1 | 8 | Done. Please.<br>しました。よろしくお願いします。 |
| | | 2 | 2 | 6 | Done. Please<br>しました、よろしくお願いします |
| @ssssseeee12341 Good afternoon<br>@ssssseeee12341 こんばんは | run3 | 5 | 0 | 2 | Aaaaaaaaaan???<br>あああああああああん??? |
| | run4 | 3 | 0 | 2 | Aaaaaaaaaan???<br>あああああああああん??? |

chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 2243–2248, Lisbon, Portugal, September 2015.

[3] M. Inoue and T. Akagi. Collecting humorous expressions from a community-based question-answering-service corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012.

[4] M. Inoue, R. Hanada, N. Furuyama, T. Irino, T. Ichinomiya, and H. Massaki. Multimodal corpus for psychotherapeutic situation. In *LREC Workshop on Multimodal Corpora for Machine Learning*, pages 18–21, Istanbul, Turkey, May 2012.

[5] M. Inoue, T. Matsuda, and S. Yokoyama. Web resource selection for dialogue system generating natural responses. In C. Stephanidis, editor, *HCI International 2011 – Posters' Extended Abstracts*, volume 173 of *Communications in Computer and Information Science*, pages 571–575. Springer Berlin Heidelberg, 2011.

[6] M. Inoue and H. Ueno. Wizard-of-Oz support using a portable dialogue corpus. In *The 5th International Workshop on Empathic Computing*, Gold Coast, Australia, Dec. 2014.

[7] L. Shang, T. Sakai, Z. Lu, H. Li, R. Higashinaka, and Y. Miyao. Overview of the ntcir-12 short text conversation task. Technical report, 2016.