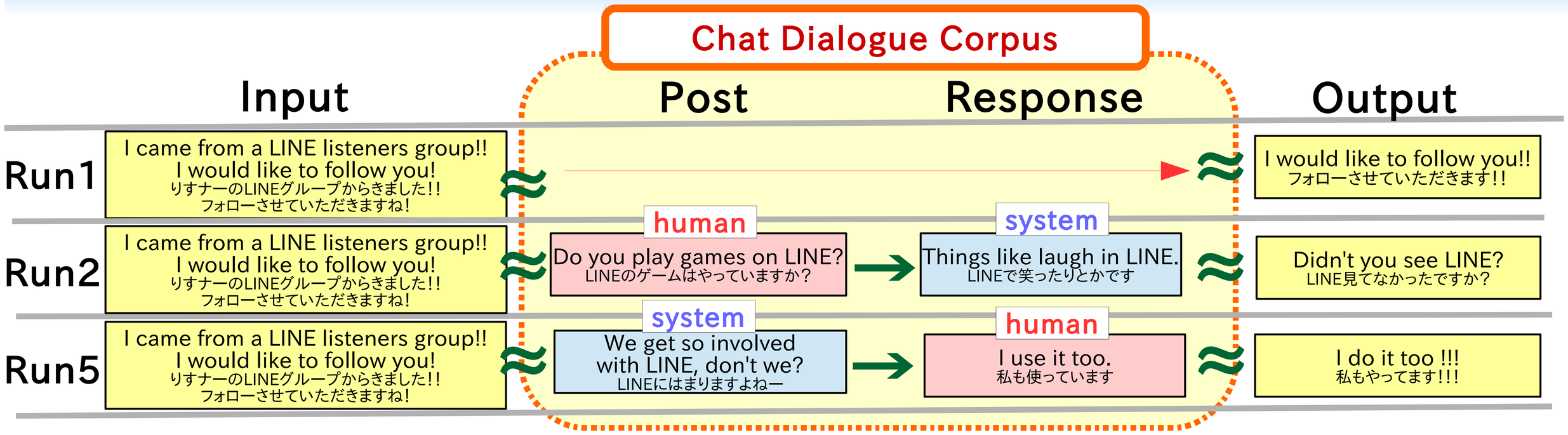
YUILA at the NTCIR-12 Short Text Challenge: Combining Twitter Data with Dialogue System Logs

Hiroshi Ueno, Takuya Yabuki, Masashi Inoue*
Yamagata University
*mi@yz.yamagata-u.ac.jp

Abstract

The YUILA team participated in the Japanese subtask of the NTCIR-12 Short Text Challenge task. We used the external dialogue log corpus. In the test run, this approach(Run2) performed far worse than the baseline(Run1). Therefore we implemented an additional experiment(Run5). The additional experiment performed much better than the first experiment but still worse than the baseline.

Method



Our approach is the use of the existing post and response relationship between texts. We used the chat dialogue corpus[1] that has been created by recording the utterance logs between users and a dialogue system.

To calculate similarities between texts in documents, we employed tfidf weighting on characters to create feature vectors, and a cosine similarity as scores.

The procedure of the proposed method (Run2) is as follows:
(1). Select most similar human's post in the corpus to input tweet.
(2). Focus on system's response to the human's post.
(3). Select most similar tweet as an output from candidate tweets to the response.

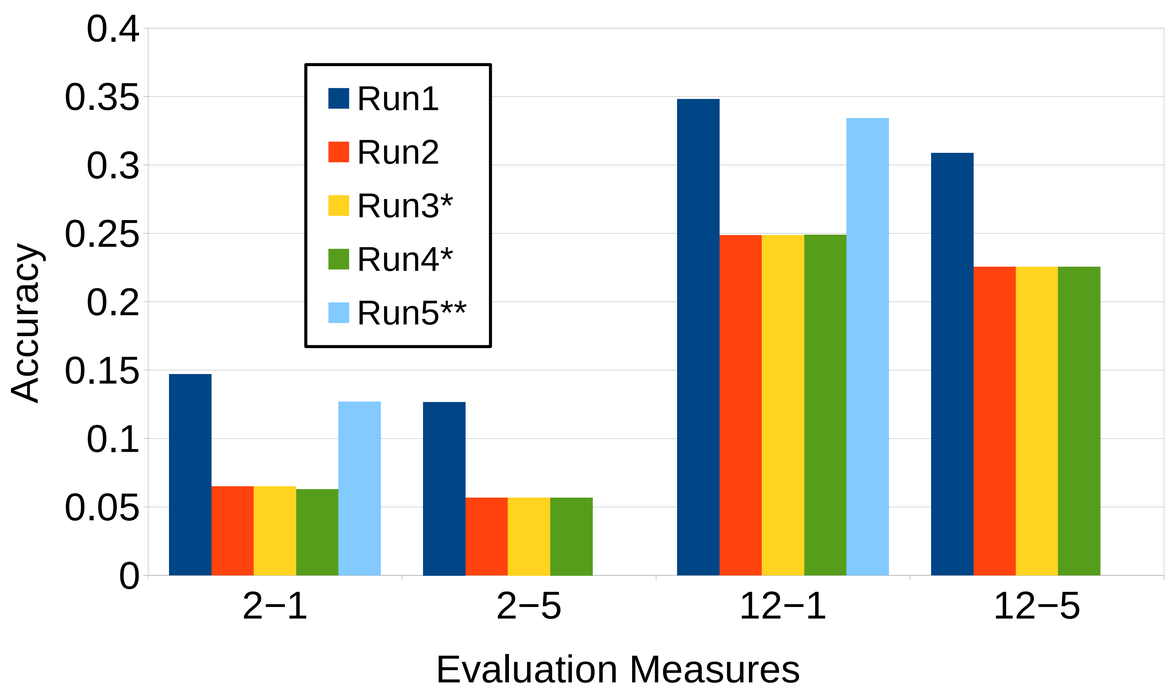
Baseline(Run1) selects most similar tweet as an output to the input tweet.

In additional run(Run5), we used human's responses instead of often irrelevant system's response. The procedure is as follows:
(1). Select most similar system's post in the corpus to input tweet.
(2). Focus on human's response to the system's post.
(3). Select most similar tweet as an output from candidate tweets to the response.

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

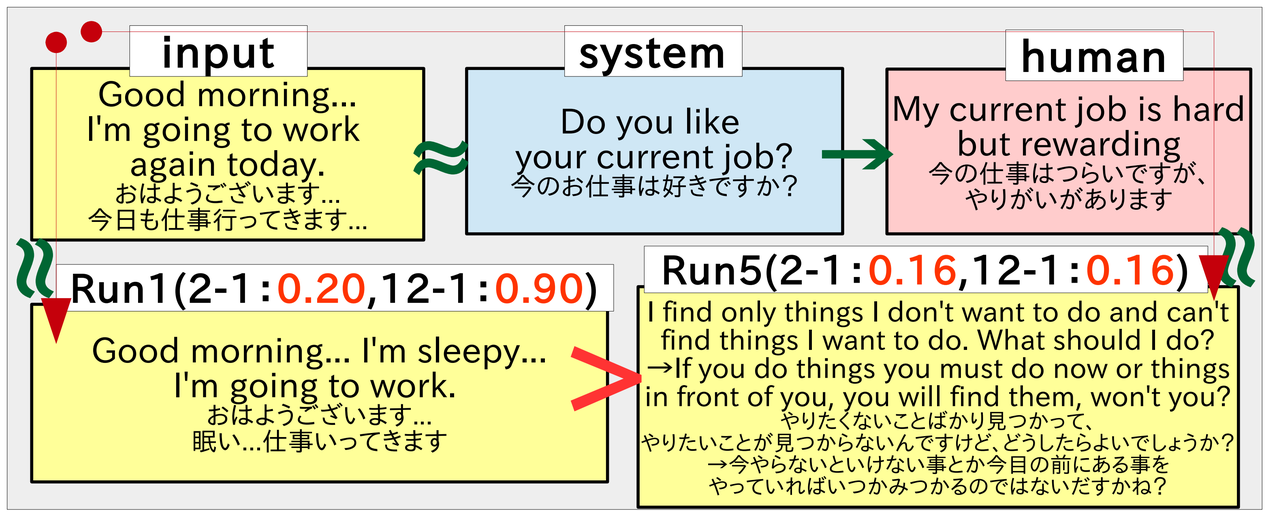
Runs

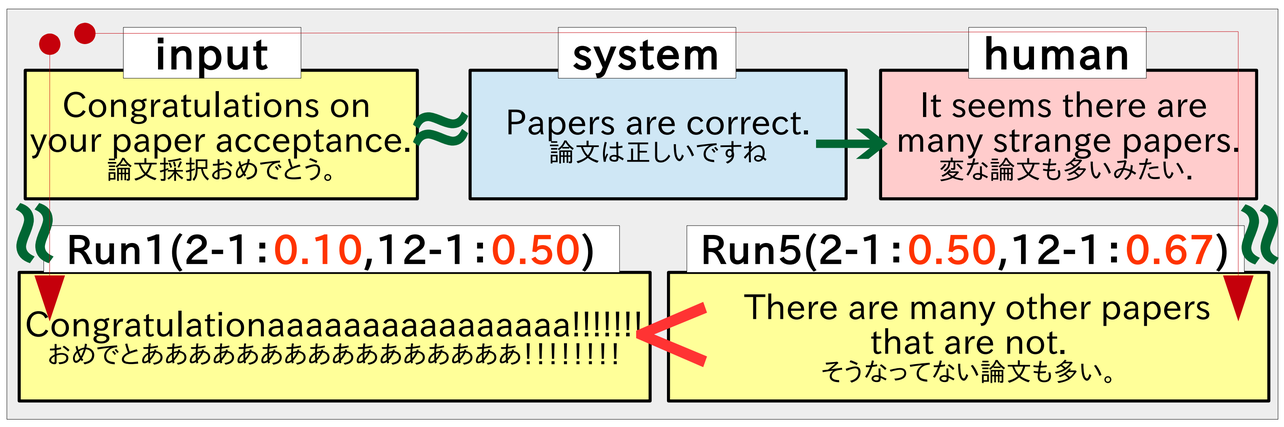
Name	Description
Run1 (BaseLine)	**Outputs are 5 most similar tweets to the input.**
Run2	**Outputs are 5 most similar tweets to the system's response in the dialogue log.**
Run3*	Outputs with rank < 5 are from Run2, and outputs with rank > 5 are from Run1.
Run4*	If there are outputs in both Run1 and Run2, they are ranked higher. Others are the same as Run3.
Run5**	**Outputs are 5 most similar tweets to the human's response in the dialogue log.**



* Results of Run3 and Run4 are almost the same as Run2 because outputs of formal runs with rank > 5 were not evaluated.
** Run5 is an additional informal run. We evaluated the result on our own by 6 evaluators using the only highest ranked output. Therefore 2-5 and 12-5 of Run5 don't exist.

Example





Discussion

Additional run improved accuracy but still worse than baseline. Although Run1that simply selects most similar text to input has a problem that may not return the answer to the question but the question to the question. Our approach may solve this problem. The failure of Run2 and Run5 indicates the semantic coherence to an input text and the dialogue coherence of utterance-response pair in using external dialogue corpus is important. Run5 has improved the dialogue coherence from Run2, but lack of the semantic coherence is a problem. For performance improvement, the investigation of features or the representation of short text and the similarity metrics are considered important.