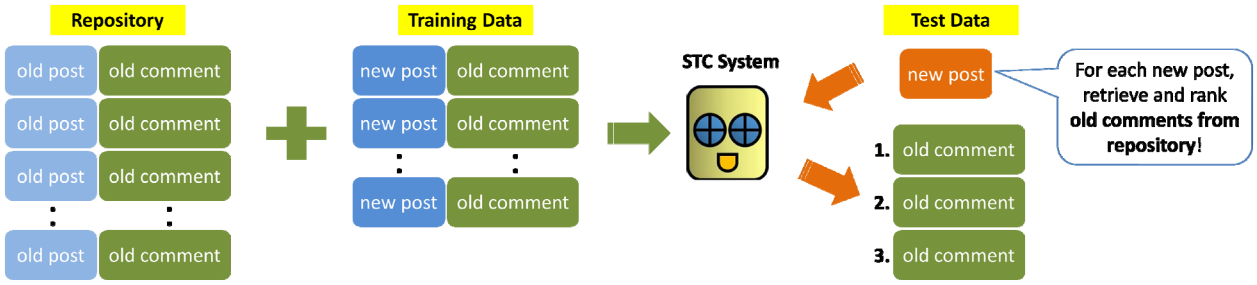
NTCIR-12 Short Text Conversation (STC) Task

Lifeng Shang♥, Tetsuya Sakai♠, Zhengdong Lu♥, Hang Li♥, Ryuichiro Higashinaka♦, Yusuke Miyao♣
Huawei♥ Waseda♠ NTT♦ NII♣ stc-org@list.waseda.jp

STC TASK DESIGN

Given a **new post**, can the system return a "good" response by retrieving a **comment** from a repository?



CHINESE SUBTASK

1. Submitted Runs

There were a total of **38** registrations, and **16** of them finally submitted **44** runs.

2. Evaluation Methods

(a) The official evaluation measures are graded relevance IR evaluation measures :***nG@1***, ***nERR@10***, and ***P+***
(b) Results from participants are pooled to perform manual annotation.

3. Chinese Test Collection

Test collection is constructed by crawling post-comment pairs from **Weibo**.

Repository	#posts	196,495
	#comments	4,637,926
	#pairs	5,648,128
Training Data	#posts	225
	#comments	6,017
	#labeled pairs	6,017
Test Data	#test topics	100

JAPANESE SUBTASK

1. Submitted Runs

There were a total of **12** registrations, and **7** of them finally submitted **25** runs.

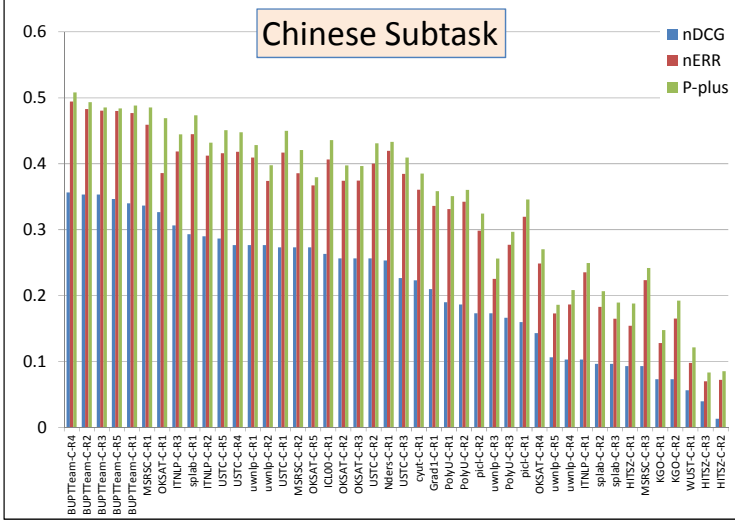
2. Evaluation Methods

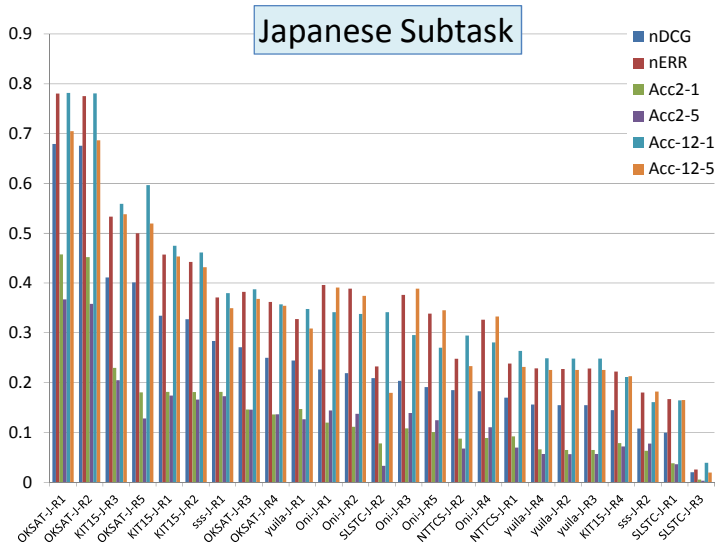
Basically the same as the Chinese subtask with the following differences:
(1) In consideration of the subjective nature of the task, the Japanese task used ten annotators to label each retrieved comment with L0, L1, or L2. For nG@1 and nERR@5, we used their average values over all annotators.
(2) In addition to ***nG@1*** and ***nERR@5***, $Acc_G@k$, which is the averaged ratio of correct labels within top-***k*** results, was used. ***G*** denotes the correct label and can either be {L1} or {L1, L2}.

3. Japanese Test Collection

Test collection is constructed by crawling tweet pairs (tweets and their replies) from **Twitter**. The training data contain 1M tweets. The test data contain 202 topics (input tweets).

Evaluation Results





Conclusions and Future work

(a) Filtering comments by using manually designed rules was simple but effective.
(b) Representing a post (or comment) by the word2vec model was helpful to perform semantic-level matching.
(c) We need to Perform more analysis on the properties of post-comment pairs from the aspects of comment length, popularity, dialogue act, and sentiment in order to learn/obtain more effective retrieval models.