CYUT-III Short Text Conversation System at NTCIR-13 STC-2 Task

Shih-Hung Wu, Wen-Feng Shih, Che-Cheng Yu,
Dept. of CSIE, Chaoyang University of Technology
168, Jifeng E.Rd. Wufeng District,
Taichung, 41349, Taiwan (R.O.C)
+886-4-23323000 ext. 4534
shwu@cyut.edu.tw, wu0fu491@gmail.com,
s10327609@gm.cyut.edu.tw

Liang-Pu Chen, and Ping-Che Yang
Institute for Information Industry,
8F, No. 133, Sec. 4, Minsheng E. Rd.
Taipei, Taiwan (R.O.C)
+886-2-66072312
eit@iii.rog.tw, maciaclark@iii.rog.tw

ABSTRACT
In this paper, we report how we build the system for Chinese subtask in NTCIR-13 Short Text Conversation (STC) shared task. We attend both retrieval-based and generation-based subtasks. Our retrieval-based system is implemented based on the Lucene search engine. The system is also used to enlarge the training data for our generation-based system, which is based on the sequence to sequence (seq2seq) model. In order to generate more comment for one given input, a feedback loop is proposed for input expansion by extracting relevant terms from the first generated comments.

Keywords
Short Text Conversation; Search Engine; sequence to sequence model; input expansion.

Team Name
CYIII

Subtasks
Short Text Conversation (Chinese)

1. INTRODUCTION
Dialogue between human and computer is a challenge task. As a simplified dialogue task, the Short Text Conversation (STC) shared task in NTCIR-13 is aimed to build a system that can retrieve or generate a comment to a given post. STC can be considered a one round dialogue. The system searches for a collection of short comments as a potential comment. The short comments was collected from social media by the task organizers [4][5].

In the retrieval-based subtask, given an input post A, the system will search the repository and find a proper comment. Since all the sentences in repository were not designed comments to the post A, it is assumed that the comments can be reused. This assumption might not be true, and the shared task is a way to verify how the sentences can be reused and make the conversation acceptable to humans.

In the generation-based subtask, given an input post A, the system will generate a proper comment. A generation model should be trained based on the given repository. This subtask is more challenging since it requires both natural language understanding and natural language generation.

At the training stage of the task, the organizer provides a training corpus, which consists of both posts and comments, and the participants create their own systems that can retrieve candidate comments. At the formal run stage, the organizer provides 100 formal test posts. In retrieval-based method the participants' systems should return 10 comments for each post. The organizers will manually check the conversation and give a relevant score from 0 to 2, where 0 is inappropriate, 1 is appropriate in some context, and 2 is appropriate in context.

Table 1 shows some sample conversation from the training corpus which was collected from a social media called SINA Weibo. These conversation is ranked 2 by the organizer.

Table 1. Sample examples of short text conversation

Post	Comment
qid_3_cid_1 京东研发体系 618 店庆指挥部在紧张工作，为京东店庆促销保驾护航！(Headquarter of jingdong research and development of shop celebration is working in the tense for escorting the jingdong shop celebration!)	repos-cmnt-2041600480 京东这个店庆太没诚意 (The shop celebration at jingdong is too insincere.)
qid_4_cid_1 移动设备改变生活，如今聚餐时确是这场景，我们成了信息的奴隶。(Mobile devices change lives, and now dinner is really like the scene, we become the slaves of information.)	repos-cmnt-2016526170 的确，移动媒体还将继续改变我们接受/处理信息和生活。(Indeed, mobile media will continue to change how we receiving / processing information and life.)
qid_14_cid_1 人生如茶，空杯以对，才有喝不完的好茶，才有装不完的欢喜和感动。(Life is like drinking tea, only with an empty cup, one can enjoy the endless good tea, and endless joy and touched moment.)	repos-cmnt-2037498400 人生如茶，静心以对 (Life is like drinking tea, face it with tranquility.)
qid_22_cid_1 很舒服的玄关设计，脱鞋穿鞋都可以坐着来~ (Very comfortable porch design, one can take off or on shoes in a sitting position ~)	repos-cmnt-2037446980 这个玄关设计的很实用、很实用，看着很舒服。`！(The design of this porch is very practical, very practical, looked very comfortable.`!)

The paper is organized as follows: We describe our methodology in section 2. Our system is shown in section 3 and 4. The result of retrieval-based system is shown and discussed in section 5. Section 6 presents the conclusion and future works.

2. Methodology

The task is about conversation; however, since the amount of candidate comments are quite large and the system has to give ten comments for each input post, we treat it as an information retrieval task in the first subtask. Our main approach to the task is to extract suitable search terms for each post and try to find comments with the search terms in the repository as the candidate comments. Our system then ranks the candidates according to the level of relevance and returns the top ten sentences as the system result for one input post.

In the second subtask, the retrieval-based system is also used to enlarge the training data for our generation-based system, which is based on the sequence to sequence (seq2seq) model.

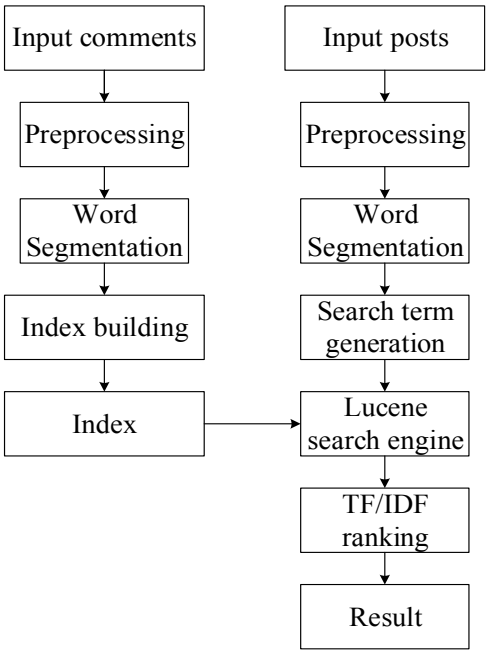

Figure 1. Retrieval-based system flowchart

3. Retrieval-based System architecture

Our system flow chart is shown in Figure 1, which consists of preprocessing module, word segmentation, index building, search engine, search term generation, and ranking module. [7]

3.1 Preprocessing

In the preprocessing module, the input sentence will be amended by filtering out some special characters. Table 2 shows some sample comments that include some special characters that will result in some problem for the next modules, such as "💝", "🐻", "☹", "(⊙_⊙)", and "🍰". Although these symbols, known as emoticons, also show the emotion in the text, we cannot process them in our system. We also normalize the punctuation symbols.

Table 2. Sample comments

Id	comments
repos-cmnt-2026963840	喜欢一切美丽事物的可以来我微博看看💝💘💓👆 (If you interested in beautiful things. You can visit my Weibo page.)
repos-cmnt-2016288310	没想到是这种结局…加油，萌🐻，未来是你们的！ (Did not expect that this outcome…keep it up, dear. The future is yours!)
repos-cmnt-2029779120	不快乐☹所以几星期没运动了。好惭愧。😑 (Not happy so that did not exercising a few weeks. Feel quite ashamed.)
repos-cmnt-2018395850	生日快乐樊老师～🍰 (Happy Birthday! Mr. Fan.)
repos-cmnt-2000215760	做错什么事了吗。(⊙_⊙) (What kind of mistake I made?)

Table 3 shows some sample posts as the input of the system. The same preprocessing will be proceeded.

Table 3. Sample posts

Id	posts
qid_0_cid_1	二战结束后，苏军士兵在德国街头强抢自行车，旁边的德国人在围观。 (After the end of World War II, Soviet soldiers rob Bicycles on the streets of Germany next to a crowd of German onlookers.)
qid_1_cid_1	夏天快到了，JMs赶紧行动起来啊~ (Summer is coming, sisters hurry up.~)
qid_2_cid_1	红会的人做慈善太累太辛苦了，听说许多人最大的心愿就是美美睡一觉 (Red cross people do charity too tired too hard, I heard that many people's biggest wish is to sleep soundly)

3.2 Word Segmentation

In any Chinese information retrieval system, the word segmentation is the first necessary step. Our system adopts an open source word segmentation tool Jieba[1]. According to the website, the tool is implemented on a variety of algorithms. Jieba can be integrated with Lucene and work well for simplified Chinese.

[1] https://github.com/fxsjy/jieba

3.3 Indexing and Search

In order to retrieve suitable candidate comments from the given comment corpus, our system uses open source tool Lucene[2] as our search engine.

Lucene was created by Doug Cutting, which is a full text search engine that can be used to build various applications [1]. Our system build on JAVA version. After filtering out stop words, the Lucene index builder can build an index for a collection of documents. Then the search engine can retrieve documents with the search terms efficiently. The default ranking mechanism is the TF/IDF ranking mechanism.

4. Generation-based System

4.1 Training Corpus Building

Our Generation-based system is built on TensorFlow Sequence to sequence (Seq2Seq) model[3] [6] with gated-unit (GRU) or long-short-term-memory (LSTM). The training set is the output of our retrieval-based system. Our retrieval-based system takes 100,000 posts as the input and retrieves 400,000 comments. We obtain a larger training set of pairs containing posts and comments. Examples of our training set is shown in Table 4.

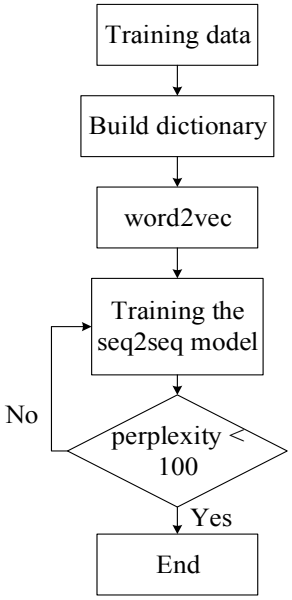

Figure 2. Training flowchart

4.2 Training the Seq2seq model

Figure 2 shows the training and test flowchart of our generation-based system. The first step is collecting all the vocabulary in the training corpus to build a dictionary. Then uses the word2vec tool to find the vector representation of each word [3]. The sentences of posts and the corresponding comments are used to train the seq2seq model. The termination criterion of training is an empirical value, perplexity equal 100. Since we do not have a validation set to find a better early stop point.

Table 4. Training data

Post	从胚胎期开始的面部特征演变过程(Evolution of facial features starting from embryonic period)
Comments	帅哥到屌丝的演变过程。 (The evolution of handsome boy to loser.)
	有正太演变成大叔的过程？(Is there a process of handsome boy becoming a big uncle?)
	《论傻 B 的演变过程》(On the Evolution Process of Stupid Baby)
	胚胎在此生根发芽，妊娠开始。(The embryo sprouts at this point and the pregnancy begins.)
Post	大学生一定要看的一分钟，它能让你奋斗一辈子(College students must watch a minute, it can make you struggle for life)
Comments	奋斗一辈子都不一定能有这么一套房(You can't have a suite like this for a lifetime.)
	一分钟说出的的话所带来的伤害，一辈子都不一定能弥补。(The damage of words in one minute, a lifetime may not be able to make up.)
	大学生，不是大学生的，学生家长一定要看！(College students, not college students, parents of students must watch!)
	一分钟就是一辈子，一辈子也就一分钟！(A minute is a lifetime, a life time is also a minute!)
Post	苹果不要只削一半好不好！你们考虑过苹果的感受嘛！(Don't just peel half of an apple! You should think about how the apple feels!)
Comments	你们到底有没有考虑过苹果的感受啊，(Have you ever thought about how an apple feels,)
	费肉！一个苹果削完，就没了一半的果肉。(What a waste! After an apple is peeled, half of it is lost.)
	单手打字多不方便你考虑过我的感受吗--咔嚓又一口苹果(one hand typing is not convenient how do you think about my feelings? -- - Snap the apple once more)
	你们考虑过胖人的感受吗。考虑过懒人的感受吗。(Have you considered the feeling of fat people? Think about how lazy people feel.)

[2] https://lucene.apache.org/

[3]https://www.tensorflow.org/tutorials/seq2seq

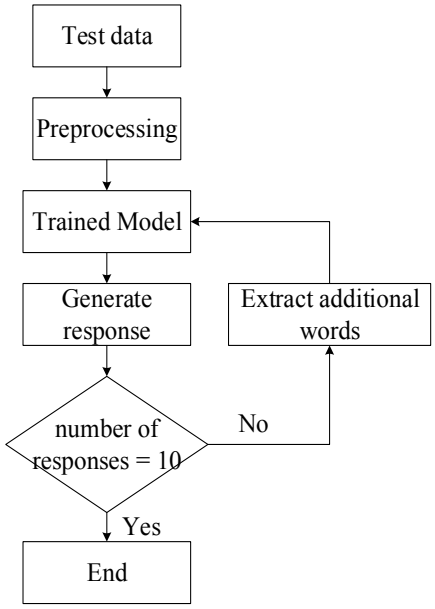

Figure 3. Generate 10 comments to the method

4.3 Generating Different Sentence by Input Expansion from a Feedback Loop

The basic seq2seq model generates only one comment to the same input. Therefore, we design a feedback loop to synthesize new inputs, so that the system can use these new inputs to get new outputs. Since the seq2seq model is very sensitive to the input, a little perturbation at an input will get quite different output.

The method is shown in Figure 3. After the first comment is generated. Our system selects words from it and with different part-of-speech (POS) tags as our feedback words, and adds them back to the original input as input expansion. The POS in use are: verbs (V), nouns (N), and adjectives (A). An example is shown in Table 5. We can see that the expanded inputs are very similar to the original post; still, it can help to generate different comments.

Table 5. The Feedback example

Post	母亲节快乐，愿天下的母亲健康漂亮 (Happy Mother's Day, Wish the world's mother be healthy and beautiful.)			
Generated comment	愿天下母亲母亲节健康 (Wish all the mothers be healthy on Mother's Day)			
The POS tags are as follows:				
愿(V) (Wish)	**天下(S)** (all)	**母亲(N)** (mothers)	**母亲节(N)** (mother's Day)	**健康(A)** (healthy)
Feedback NVA post	母亲节快乐，愿天下的母亲健康漂亮 愿天下母亲母亲节健康			
Generated comment	母亲节愿天下母亲节健康			
Feedback NV post	母亲节快乐，愿天下的母亲健康漂亮 愿天下母亲母亲节			
Generated comment	愿天下母亲节健康			
Feedback NA post	母亲节快乐，愿天下的母亲健康漂亮 天下母亲母亲节健康			
Generated comment	母亲节健康			
Feedback VA	母亲节快乐，愿天下的母亲健康漂亮 愿健康			
Generated comment	母亲节快乐，愿天下的母亲健康漂亮 愿健康			
Feedback N post	母亲节快乐，愿天下的母亲健康漂亮 天下母亲母亲节			
Generated comment	母亲节愿母亲母亲节健康			
Feedback A post	母亲节快乐，愿天下的母亲健康漂亮 健康			
Generated comment	母亲节，愿母亲健康的健康			
Feedback V post	母亲节快乐，愿天下的母亲健康漂亮 愿			
Generated comment	爱的愿母亲的母亲健康平安。			

5. Experiment Result

Table 6 shows the 5 different settings of our 5 formal runs. In the five formal runs, we tested the results of using different feedback words with different POS combination. Also, we test use the LSTM vs. GRU.

Table 6. Setting of our 5 formal runs

Run	Feedback	RNN
CYIII-C-G 1	N,V,A	GRU
CYIII-C-G 2	N,V	GRU
CYIII-C-G 3	N	GRU
CYIII-C-G 4	V	GRU
CYIII-C-G 5	N,V,A	LSTM

5.1 Formal run results

The result of the formal run of the Chinese subtask STC task is shown in Table 7. The first row is the result of the best team, and the following rows are result of our team.

Table 7. Formal run results in STC task

Run	Mean nG@1	Mean P+	Mean nERR@10
Best results in formal run	**0.5867**	**0.6670**	**0.7095**
CYIII-C-R1	0.4262	0.5332	0.5668
CYIII-C-G1	0.1213	0.1684	0.1771
CYIII-C-G2	0.1163	0.1577	0.1662
CYIII-C-G3	0.092	0.1234	0.1366
CYIII-C-G4	0.107	0.1462	0.1536
CYIII-C-G5	0.0783	0.1101	0.115

5.2 Formal run Evaluation Measures

The evaluation metrics of the STC is based on the three metrics: mean nG@1, mean P+, and mean nERR@10.

5.2.1 nG@1

The nG@1 is an IR metrics that takes the ranking into account. A system can get the highest score if the system ranks the retrieved document perfectly. According to the organizer, the nG@1 is defined as:

$$nG@1 = \frac{g(1)}{g^*(1)}, \quad (1)$$

which only consideres the top 1 answer. And the only possible *g()* values are 0, 1/3, and 1. Corresponding to the 0, 1, and 2 manually labelled score, $g^*()$ represents the perfect result.

5.2.2 nERR@10

Expected Reciprocal Rank (ERR) [1] is also used. The score of a retrieved document is defined as $p(r) = \frac{g(r)}{2^H}$, where H denotes the highest relevance level, in the STC task is 2. Therefore, if the retrieved document is L2-relevant, $p(r) = 3/4$; if the retrieved document is L1-relevant, $p(r) = 1/4$;

if the retrieved document is L0-relevant, $p(r) = 0$. Normalized ERR at a cutoff 10 is as follows:

$$nERR@10 = \frac{\sum_{r=1}^{10} Pr_{ERR}(r)\left(\frac{1}{r}\right)}{\sum_{r=1}^{10} Pr^*_{ERR}(r)\left(\frac{1}{r}\right)} \quad (2)$$

5.2.3 P+

The P^+ metric, similar to ERR, was proposed in AIRS 2006 [2]. Given a ranking list, r_p represents the rth document in the list. Just as the weighting in ERR is (1/r), in P+ the weighting is BR(r):

$$BR(\mathrm{r}) = \frac{\sum_{k=1}^{r} I(k) + \sum_{k=1}^{r} g(k)}{r + \sum_{k=1}^{r} g^*(k)} \quad (3)$$

where g and g* are defined as in nG@1.

5.3 Discussions

The result of the formal runs show the weaknesses of our retrieval-based system. The first one is that we treat the task as an information retrieval task and only try to find related sentence as the comment. This is not a comprehensive approach to the task. A better understanding is needed for a conversation.

About the word segmentation, we adopt the Jieba tool and it is much better than Jeseg that we used in STC-1. But the lexicon of the word segmentation system is not rich enough to cover the vocabulary used in the corpus, especially the lack of specific terminology. More lexicon is needed; therefore, we have to collect new words from various sources.

Ranking of the retrieved comments is also an issue. Since our system tends to retrieve related sentence, the candidate sentences were shown no compassion at all, and this makes the system output a poor conversation.

5.4 System Analysis

There are some test posts that our system gives such good comment as the following example in Table 8. We can find that in these cases, our system searches for the sentences with the terms appearing in the posts and finds L2-relevant comments. Because similar terms can construct similar sentences, and sometimes, in a good conversation, to confirm what is said is a good comment.

Table 8. Sample cases that our system can output more L2-relevant comments

Posts	Retrieval Comments
test-post-10080 吃素第一天，坚持住，崔朵拉。 (Vegetarian first day, hold on, Tridola.)	repos-cmnt-2021212720 坚持住，一定要坚持住。 (Hold it on, keep hold it on.)
test-post-10120 单车骑行 4 年穷游 30 国，小伙变大叔。 (After cycling 4 years and touring 30 countries, a small boy becoming a big uncle.)	repos-cmnt-2033159820 4 年了，变成熟了，但更有魅力了。 (After four years, become mature and more charm.)
test-post-10300 今晚又可以在海边发呆了，舒畅啊。 (Tonight can be in a daze at the seaside, comfortable ah.)	repos-cmnt-2034755820 噢噢噢我最喜欢在海边发呆了 (Oh, oh, oh, I love being in a daze at the seaside.)

There are some test posts that our system cannot give good comment such as the following example in Table 9. Again, our system searches for the sentences with the terms appearing in the posts but finds L0-relevant comments. Since these terms mean differently in the posts and in the comments, the comment makes little sense in the conversation.

Table 9. Sample case that our system can only output L0-relevant comments

Posts	Retrieval Comments
test-post-10680 我眼睛的度数好像又增加了，完全看不见你的智商了 (The degree of my eyes seems to have increased again, completely can't see your IQ)	repos-cmnt-2005954660 智商和眼睛度数成正比。 (IQ is proportional to the degree of the eye.)
test-post-10730 喜欢宫崎骏的童鞋千万别错过，快来看看吧。 (don't miss Hayao Miyazaki's children's shoes, come and see it.)	repos-cmnt-2034306860 喜欢玩俄罗斯方块儿的童鞋快来看。好有感。 (Come to see the children's shoes with Tetris. Good feeling.)

We find that sentences with similar meaning can be a good comment although it is not the only way to make a good conversation. And we also find that the sentences which look similar to the post but have different meanings are not good comments. Our generation-based system usually generates sentences with grammar errors, which do not meet the requirement of fluent. Table 10 shows the better sentences generated by our system. It can generate totally different comments to the same post from the comments retrieved by the retrieval-based system. But the comments are still not self-sufficient, since the quality and quantity of training set are not sufficient.

Table 10. Better cases of our generation-based system

Post	test-post-10190 鼓励种转基因作物的种子公司都是玩弄各国政府与各国农业部还有农民的骗子 (Seed companies that encourage genetically engineered crops are crooks who manipulate governments and ministries of agriculture and farmers)
Generated comment	这谎言是谎言，谎言是谎言，还是要被谎言！(This lie is a lie, a lie is a lie, or be a lie!)
Post	test-post-10230 好喜欢这种手绘的画啊，小柚子加油 (Like this hand-painted painting, go go little grapefruit)
Generated comment	好喜欢画的画的画的好温馨啊。(I like the paintings with warm feeling, ah.)
Post	test-post-10250 和家人一起喝喝茶，聊聊天，也是一种生活的乐趣 (Drinking tea and chatting with family members is also a fun way to live.)
Generated comment	喝茶和家人一起喝茶，也是一种幸福的事情！(Drink tea with family members is also a happy thing!)
Post	test-post-10580 听说最近风很大，大家一定要多穿，不要被风吹跑噢 (I heard that the wind is very strong recently, we must wear more, do not be blown away by the wind OH)
Generated comment	今天出门也要注意了，要注意身体哦。(Today go out also pay attention to the body oh.)
Post	test-post-10770 想支持一下国产手机 (Want to support the domestic mobile phone)
Generated comment	就支持一下。(Support it.)
Post	test-post-10900 又老了一岁，感谢我们组，我很高兴 (One year older, thanks to our group, I am very happy)
Generated comment	生日快乐！老了！老！(Happy Birthday! Getting old! old!)

6. Conclusions and Future Works

In this paper, we report how we build the system for Chinese subtask in STC shared task. In retrieval-based method, our system can find the most related comments for a given input post.

There are some future works to improve the system. The first one is to enlarge the lexicon of the word segmentation system. Many new terms appear in the test post; in addition to human name and place name, there are new Internet slams. These out-of-vocabulary terms do decrease the performance of our system, and we need to collect them from the Internet regularly.

The types of posts are also need to explore. There are many posts involved in social events and entertainments. The conversation is not just question answering, it also contains discussions on subjects in many cases. A proper comment in these cases should be a confirmation on an opinion or give a counter question.

The ranking of the comments can be a separate issue. Ranking by learning is a promising approach. Since it is not a retrieval task or a question answering task, the ranking reason should be more subtle than traditional TF/IDF.

The sentence generated by the generation-based system are totally different from comments retrieved by the retrieval-based system. Even though the training set is provided by the retrieval-based system. To generate better comments, the system need better quality and larger quantity of training set.

According to our observation of the corpus, we find some dialog principles suitable for certain comments. For example, if someone feels sad and says “today is not my day”, then some encouraging words are suitable in this case. So, recognizing the different situations might be a meta condition for conversation generation.

7. Acknowledgment

This study is conducted under the "Online and Offline integrated Smart Commerce Platform (4/4)" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China. This study is also supported by the Ministry of Science under the grant numbers MOST106-2221-E-324-021-MY2.

8. REFERENCES

[1] O. Chapelle, S. Ji, C.Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: Metrics and algorithms. Information Retrieval, 14(6):572-592, 2011.

[2] T. Sakai. Bootstrap-based comparisons of IR metrics for finding one relevant document. In AIRS 2006 (LNCS 4182), pages 374–389, 2006.

[3] Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." arXiv preprint arXiv:1405.4053 (2014).

[4] Lifeng Shang, Tetsuya Sakai, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao, Yuki Arase, Masako Nomoto. Overview of the NTCIR-13 Short Text Conversation Task, in Proceedings of NTCIR-13, 2017.

[5] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao. Overview of the NTCIR-12 Short Text Conversation Task, in Proceedings of NTCIR-12, 2016.

[6] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[7] Shih-Hung Wu, Wen-Feng Shih, Liang-Pu Chen and Ping-Che Yang, CYUT Short Text Conversation System for NTCIR-12 STC, in Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, June 7-10, 2016, Tokyo Japan, pp.541-546.