CYUT Short Text Conversation System for NTCIR-12 STC

Shih-Hung Wu [1], Wen-Feng Shih [2], Liang-Pu Chen[3], and Ping-Che Yang[4]
[1, 2] Dept. of CSIE, Chaoyang University of Technology, Taichung, Taiwan (R.O.C)
[3,4]Institute for Information Industry, Taipei, Taiwan (R.O.C)
shwu@cyut.edu.tw[1], wu0fu491@gmail.com[2], eit@iii.org.tw [3], maciaclark@iii.org.tw[4]

Abstract

In this paper, we report how we build the system for Chinese subtask in NTCIR12 Short Text Conversation (STC) shared task. Our approach is to find the most related sentences for a given input sentence. The system is implemented based on the Lucene search engine. The result shows that our system can deal with the conversation that involves related sentences.

Keywords: Short Text Conversation; Search Engine; text similarity;

1. Introduction

Dialogue between human and computer is a challenge task. A system that can give a proper response to any input sentence requires various kinds of knowledge, such as natural language understanding, language inference and common sense. Although there are some commercial systems in real world for some specific domains, it requires lots of data and research to achieve a better dialogue system.

As a simplified dialogue task, the Short Text Conversation (STC) shared task in NTCIR 12 is aimed to build a system that can give a response to a given post [7]. STC can be regarded as a one round dialogue. The system searches for a collection of short comments as a potential response. The short comments was collected from social media by the task organizers.

Given the input post A, the system will search the repository and find a proper response. Since all the sentences in repository were not designed responses to the post A, it is assumed that the responses 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 human.

At the training stage of the task, the organizer announced a training corpus, which consists of both posts and response comments, the participants can create their own systems that can retrieve candidate responses. At the formal run stage, the organizer announced 100 formal test posts. The participants' systems should return 10 response comments for each post. The organizers will manually check the conversation and give a relevant score from 0 to 2, where 0 is for not proper or L0-relevant, 1 is for marginal in the context or L1-relevant, and 2 is for a proper response or L2-relevant.

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	Response Comment
train-post-10010 昨天晚上程序挂掉了。(Last night, my process died.)	repos-cmnt-1025051760 (2) 亲身经历过的人默默无语路过。。。。。(As a person with the same experience, I cannot say anything.)
train-post-10390 只要有信心，人永远不会挫败。(With confidence, one will never be defeated.)	repos-cmnt-1027397270 (2) 自己要给自己信心。(You should have confidence.)
train-post-11120 当我一个人想太多的时候，神经就会变得敏感。(When I thing intensively, I become to sensitive.)	repos-cmnt-1034226450 (2) 为此我要变得不敏感 (Therefore, I shall become less sensitive.)
train-post-12000 又一天结束 下班回家 7-Eleven 便当还是永和豆浆？(On my way home, you want me buy lunch a boxed meal or soy milk?)	repos-cmnt-1012480480 (2) 永和豆浆上榜 (soy milk is on the list)

The paper is organized as follows: The methodology is briefly described in section 2. We describe our system in section 3. The result is shown and discussed in section 4. Section 5 gives 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 responses for each input post, we treat it as an information retrieval task. 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 responses. Our system then ranks the candidates according to the level or relevance and returns the top ten sentences as the system result for one input post.

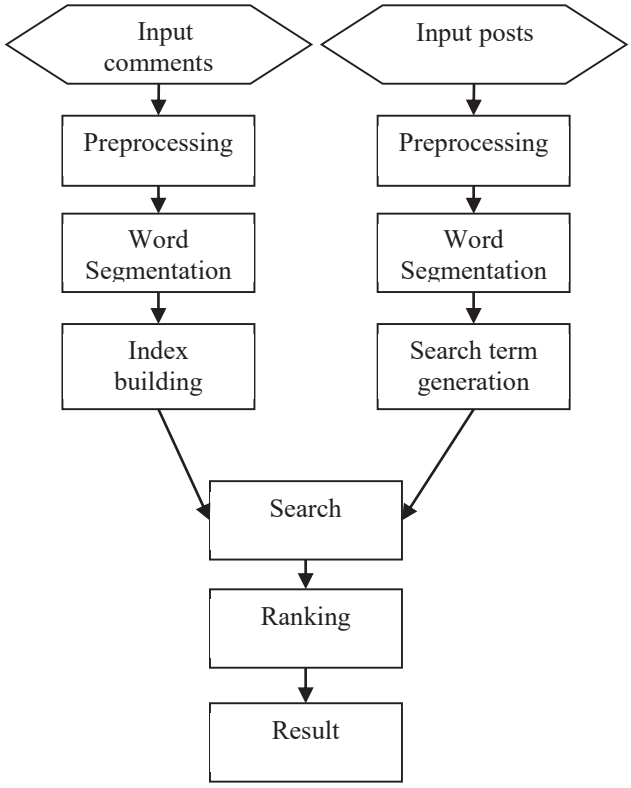

Figure 1. System flowchart

3. 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.

3.1 Preprocessing

In the preprocessing module, our system will modify the input sentence by filtering out some characters. Table 2 shows some sample comments that contain characters that will course trouble for the next modules, such as " ", "💘", "💯", "🙆", and "～". Although these symbols, known as emoticans, 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	comment
repos-cmnt-1000000010	太危险了 (It is too dangerous)
repos-cmnt-1000000020	不会姓鲍吧 (Shouldn't the family name be Bau?)
repos-cmnt-1000000030	薄案体现了中央反腐决心，坚决拥护党中央，打老虎不手软! (The Bo case reflects the determination of central government on anti-corruption. Beat the tigers without hesitation!)
repos-cmnt-1000009420	祛痘/除痘印💘两周见效💯有意者可联系我🙆 (Remove Acne / Acne mark within two weeks, 100% effective, contact me)
repos-cmnt-1000008920	坑老子啊～～～ (Do not defraud me)

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

Table 3. Sample posts

Id	post
train-post-10010	昨天晚上程序挂掉了。 (Yesterday evening the program terminated unexpectedly.)
train-post-10020	今晚大餐:豪尚豪牛排 (Special dinner: steak Hao Shang Hao)
train-post-10030	准备出发享受阳光沙滩 (Ready to enjoy the sunny beaches)
train-post-10040	到家了，感觉好温馨啊 (Get home, feeling warm)
train-post-10050	今晚主打私藏的黄酒。 (Tonight the main course is the collected rice wine.)

3.2 Word Segmentation

Word segmentation is the first step for any Chinese information retrieval system. Our system adopts an open source word segmentation tool Jcseg [2], the version number is 1.9.4. According to the website, the tool is an implementation of mmseg algorithm [2], the test accuracy is 98.41%. Jcseg can be integrated with Lucene and work well for simplified Chinese.

3.3 Indexing and Search

To retrieve candidate response from the given comment corpus, our system uses open source tool Lucene [3] 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] We use the JAVA version to build our system. After filtering out stop word, 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. Experiment Result

4.1 Formal run results

The result of the formal run of the Chinese subtask STC task is shown in Table 3.

Table 3. Formal run results in STC task

Run	Mean nDCG@1	Mean P+	Mean nERR@10
Best results in formal run	**0.3567**	**0.5082**	**0.4945**
cyut-C-R1	0.2233	0.3851	0.3608

4.2 Formal run Evaluation Measures

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

4.2.1 nDCG@1

The nDCG 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 considered 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.

4.2.2 nERR@10

Expected Reciprocal Rank (ERR) [3] 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:

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

4.2.3 P+

The P^+ metric, similar to ERR, was proposed in AIRS 2006 [4]. Given a ranking list, let r_p be 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 nDCG.

4.3 Discussions

After the formal run, we find that there are some drawbacks in our system. The first one is that we treat the task as an information retrieval task and only try to find related sentence as the response. This is not a comprehensive approach to the task. A better understanding is needed for a conversation.

A technical problem that we met is the word segmentation. 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 show no compassion at all, and this makes the system output not as a good conversation.

4.4 System Analysis

There are some test posts that our system gives good response. Such as the following example in Table 4. We can find that in these cases, our system searches the sentences with most terms in the posts and finds L2-relevant responses. Because similar terms can construct similar sentences, and sometimes, in a good conversation, to confirm what is said is a good response.

Table4. Sample cases that our system can output more L2-relevant responses

Post	Response Comment
test-post-10010 远离城市，才得安宁。 (Far away from the city, then quietness could be found)	repos-cmnt-1038680270 远离城市的喧嚣，美丽的宁静 (Far away from the noisy city, beautiful quietness)
test-post-10010 远离城市，才得安宁。 (Far away from the city, then quietness could be found)	repos-cmnt-1037175540 追逐内心的安宁，那么便是远离了安宁。 (Looking for the quietness inside, is drifting from quietness.)
test-post-10300 保持乐观的生活态度。 (Keep an optimistic outlook on life.)	repos-cmnt-1018944480 乐观的生活态度最重要 (Optimistic attitude towards life is most important)

There are some test posts that our system cannot give good response. Such as the following example in Table 5. Again, our system searches the sentences with most terms in the posts but finds L0-relevant responses. Since these terms means differently in the posts and in the comments, the response makes little sense in the conversation.

Table5. Sample case that our system can only output more L0-relevant responses

Post	Response Comment
test-post-10860 创业就像下棋，你有对手，但最大的对手永远都是你自己。 (Being an entrepreneur is like playing chess, you have opponents, but the biggest rival will always be yourself.)	repos-cmnt-1001791750 创业者，是通过竞争对手了解自己的 (Entrepreneurs learn themselves by understanding their competitors)
test-post-10960 如果你希望现在与过去不同，请研究过去。 (If you want to be different from the past, study the past.)	repos-cmnt-1001458230 现在的左派，已然与过去的左派不同了。 (The current left wing, is different already from the left wing in the past.)

We find that sentences with similar meaning can be a good response, although not the only way to make a good conversation. And we also find that, the sentences that looked similar to a post but have different meanings are not good responses.

4.5 The relationship between post and responses

By observing the training set, we find that there some communication types between the posts and their responses. Here we list 7 major types: sympathy, confirmation, reply to information, sarcasm, Nonsense, ask back, and envy. Examples are given in Table 6.

1. Sympathy
   The response shows feelings of pity and sorrow for the misfortune in the post.
2. Confirmation
   The response is confirming the opinion or information in the post.
3. Reply to information
   Whey post is a query for some kind of information, the response should be a reply to the question.
4. Sarcasm
   Some response is laughing or deny the idea in the post.
5. Nonsense
   Nonsense some time is a kind of humor; however, literally the response is nonsense.
6. Ask back
   Some responses are questions that ask back the same thing in the posts.
7. Envy
   Some response shows the feeling of envy to the post.

Table6. Sample case that relationship between post and responses in the training set

Type	Post and response example	
1	**Post**	train-post-10010 昨天晚上程序挂掉了。 (Yesterday evening the program terminated unexpectedly.)
	Response Comment	repos-cmnt-1025051760 亲身经历过的人默默无语路过。。。。。(As a person with the same experience, I cannot say anything.)
2	**Post**	train-post-10110 常打羽毛球，预防颈椎病 (By playing badminton regularly, one can prevent the cervical disease)
	Response Comment	repos-cmnt-1015755620 不错不错，值得推广 (good idea, worthy of promotion)
3	**Post**	train-post-10310 周末去三亚玩，有啥推荐？！ (Visisting Sanya this weekend, so what's recommended? !)
	Response Comment	repos-cmnt-1037288230 亚农湾，热带雨林(小朋友认植物)，温泉(老人喜欢) (Yalong Bay, rainforest (for kids to learn plant), hot springs (good for elderly))
4	**Post**	train-post-11200 夏天有了这个帽子，妈妈再也不担心我被晒黑了！ (With this summer hat, my mother no longer worried that I will be tan!)
	Response Comment	repos-cmnt-1026413870 够黑了。。。 (Tan enough)
5	**Post**	train-post-10260 想去武当山 有想同游的么？ (Anybody want to travel Wudang with me?)
	Response Comment	repos-cmnt-1044036510 哈哈 我也好想请你去 KTV (Haha I really want to ask you to go KTV)
6	**Post**	train-post-11230 最幸福的事，莫过于有个人总是喜欢逗你开心。。 (Nothing is more happier than someone always loves to make you happy. .)
	Response Comment	repos-cmnt-1018063370 逗你开心幸福吗？ (Is the one who making you happy happy?)

7	**Post**	train-post-11440 忍不了了，今晚大吃羊肉，去火的事情明天交给中医~ (Can't put up tonight for eating mutton, the consequence left to the Traditional Chinese Medicine doctors tomorrow ~)
	Response Comment	repos-cmnt-1029818590 看得俺口水嘀嗒的。。 (Make my saliva ticks. .)

We also find examples in the test set. As shown in table 7.

Table7. Sample case that relationship between test-post and responses in the test set

Type	Post and response example	
1	**Post**	test-post-10050 一大早上，各种打哈欠的，有木有
	Response Comment	repos-cmnt-1029825440 惨了，我经常一起床就打哈欠，而且经常被传染打哈欠.....
2	**Post**	test-post-10010 远离城市，才得安宁。
	Response Comment	repos-cmnt-1044486650 远离城市的渲嚣，宁静致远
3	**Post**	test-post-10310 可穿戴设备会对我们的生活造成什么样的影响？
	Response Comment	repos-cmnt-1007666350 木有影响，照常生活
4	**Post**	test-post-10320 天天就知道耍猴，这次被猴耍了吧！
	Response Comment	repos-cmnt-1035749290 哈哈，被耍了
5	**Post**	test-post-10060 今天农历正月十一，是俺的出生日，求祝福！
	Response Comment	repos-cmnt-1023807350 宗教信仰自由，哈哈，希望十一郎也修炼成佛，那可真是国家之幸。
6	**Post**	test-post-10000 到青岛，一个比印象更好的城市。到了才想起火车票忘给钱了
	Response Comment	repos-cmnt-1040207300 你们在青岛？
7	**Post**	test-post-10340 重庆小雨，三五人围坐火锅！香吃无语。
	Response Comment	repos-cmnt-1011999470 想吃啊，流口水了。

4.6 Analogy Approach

According to the types that we observed, we find a new way based on the analogy word embedding to get the proper responses. A post A and a response B in the training set can be coupled to find the best response D for a test post C by the analogy A:B = C:D.

We adopted the wor2vec toolkit by Mikolovs [5] to conduct the experience. The toolkit in python can be found atGitHub[6]. Some result is shown in Table 8.

Table8. The experimental results by analogy approach

1	**Post A**	train-post-10010 昨天晚上程序挂掉了。 (Yesterday evening the program terminated unexpectedly.)
	Response Comment B	repos-cmnt-1025051760 亲身经历过的人默默无语路过。。。。。(As a person with the same experience, I cannot say anything.)
	Post C	test-post-10000 到青岛，一个比印象更好的城市。到了才想起火车票忘给钱了
	Response Comment D	repos-cmnt-1004656480 吐血了。。。。
2	**Post A**	train-post-10110 常打羽毛球，预防颈椎病
	Response Comment B	repos-cmnt-1000118190 好！
	Post C	test-post-10010 远离城市，才得安宁。
	Response Comment D	repos-cmnt-1004656480 这个好！
3	**Post A**	train-post-10050 今晚主打私藏的黄酒。
	Response Comment B	repos-cmnt-1045083900 黄酒应该尝尝镇平的。
	Post C	test-post-10000 到青岛，一个比印象更好的城市。到了才想起火车票忘给钱了

	Response Comment D	repos-cmnt-1020438170 下次回哈尔滨要去看看。
4	**Post A**	train-post-11200 夏天有了这个帽子，妈妈再也不担心我被晒黑了！
	Response Comment B	repos-cmnt-1026413870 够黑了。。。
	Post C	test-post-10030 晚上加餐了，在宾馆旁边喝羊汤
	Response Comment D	repos-cmnt-1045759290 这个味道可恶心了。。。。。
5	**Post A**	train-post-10400 打球被人指甲划了表示很受伤…
	Response Comment B	repos-cmnt-1024991800 肉灵芝表示很受伤，无辜被人扣上这黄色帽子
	Post C	test-post-10130 终于看到蓝天啦！和坏天气 saybye！
	Response Comment D	repos-cmnt-1005597660 可不想成功的人就永远不会成功
6	**Post A**	train-post-10070 五一，你准备去哪儿？
	Response Comment B	repos-cmnt-1003988340 有过五一吗？
	Post C	test-post-10050 一大早上，各种打哈欠的，有木有
	Response Comment D	repos-cmnt-1000295940 是吗？？？
7	**Post A**	train-post-10230 睡前，想吃点儿甜点...
	Response Comment B	repos-cmnt-1021039530 各种想吃
	Post C	test-post-10340 重庆小雨，三五人围坐火锅！香吃无语。
	Response Comment D	repos-cmnt-1000967570 好想吃啊

The results shown in Table 8 have some interesting characteristics that are quite different from the results of our system in the official runs. Since the analogy approach do not rely on the keyword search technology, the response contains no keywords of the post in most cases. And still, the responses make sense in these cases. We believes that this approach can be used in broader cases that keyword search approach does not work well.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we report how we build the system for Chinese subtask in STC shared task. 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 type 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 response in these cases should be a confirmation on the opinion or give a counter question.

The ranking of the responses 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.

REFERENCES

[1] Apache Lucene, https://lucene.apache.org/

[2] Chinese Word Segmentation toll jcseg, https://code.google.com/p/jcseg/

[3] 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.

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

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

[6] sentence2vec - Tools for mapping a sentence with arbitrary length to vector space. https://github.com/klb3713/sentence2vec

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