WUST System at NTCIR-12 Short Text Conversation Task

Maofu Liu, Yifan Guo, Yang Wu, Limin Wang
College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, P.R. China
liumaofu@wust.edu.cn, 498966594@qq.com, wuyang0329@foxmail.com, smile_wlm@163.com

Han Ren
College of Computer Science, Hubei University of Technology, Wuhan 430068, P.R. China
hanren@whu.edu.cn

ABSTRACT

Our WUST team has participated in the Chinese subtask of the NTCIR-12 STC (Short Text Conversation) Task. This paper describes our approach to the STC and discusses the official results of our system. Our system constructs the model to find the appropriate comments for the query derived from the given post. In our system, we hold the hypothesis that the relevant posts tend to have the common comments. Given the query *q*, the topic words firstly are extracted from *q*, and the initial set of post-comment pairs retrieved, and then used to match and rank to produce the final ranked list. The core of the system is to calculate the similarity between the responses and the given query *q*. The experimental results using the NTCIREVAL tool suggest that our system should be improved by combining with related knowledge and features.

Keywords

Short text conversation, Information retrieval, Vector space model

1. INTRODUCTION

It is well known that the human-computer conversation is one of the difficult artificial intelligence problems, which aims to understand the meaning and intention of the texts. The progress of this problem is unfortunately still quite limited. At NTCIR-12, the STC task is a much simplified version of this problem: One round of conversation formed by two short texts, with the former being an initial post from users and the latter being a comment given by the computer [1]. The STC task at NTCIR-12 consists of two subtasks, i.e. one is Chinese and the other is Japanese. The main difference between the two subtasks lies in sources and languages of test collections [4]. We only concerned with the Chinese subtask in this paper.

Sina Weibo has become an indispensable communication platform, on which millions of users publish their short messages and the other users can comment on a published post [2]. Sina Weibo complies with the word limit of 140 Chinese characters in one post or comment, and provides vast real-world instances in Chinese, forming a natural dataset of one-round conversation. The comments to a post can be of rather flexible forms and diverse topics. As a result, instead of generating a response to a post, we pick the most appropriate one from the dataset for the given post on STC task.

The STC is defined as an IR (Information Retrieval) problem to search for the appropriate comments matching the given query *q*, derived from the given post, from the dataset. That is to say, given a query *q*, the system with a reasonable model and a large-scale of candidates can automatically produce fairly natural and appropriate responses. From the theoretical perspective, this work tries to verify that the relevant posts tend to share similar content and common comments.

In this paper, we propose the model to estimate whether the post and the given query *q* are relevant. We adopt the framework to extract topic-words from the given query q, to retrieval candidate responses, and then match and rank the responses to produce the final ranked list. The core of the system is to evaluate the relevance. And then we use a simple VSM (Vector Space Model) for measuring the similarity between the given *q* and the post.

The remainder of this paper is organized as follows. Section 2 discusses our system architecture in details. Section 3 describes our evaluation results of the formal run on the test collection of STC task. Finally, we conclude our paper in section 4.

2. SYSTEM DESCRIPTION

Our system includes three main stages, i.e. data preprocessing, matching and ranking. Figure 1 illustrates our system architecture in detail.

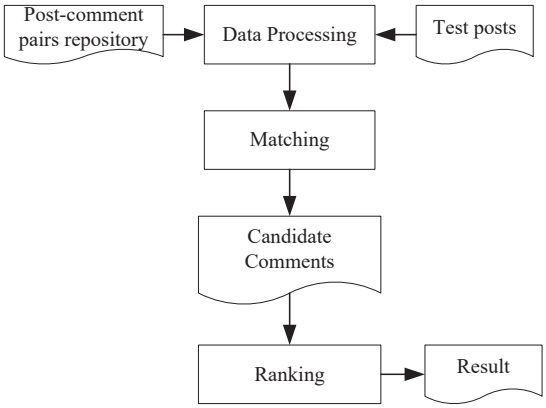

Figure 1. System overview

2.1 Data preprocessing

This stage includes two parts. In part one, the system processes the post-comment pair repository and builds the inverted index table of the posts and words separately. The inverted index table of words records the frequencies and location indexing where the post contains the word. Table 1 shows the number of the post-comment pair repository and test dataset for Chinese STC task.

Table 1. Statistics of the dataset for Chinese STC task

Retrieval Repository	posts	196,495
	post-comment pairs	5,648,128
Test Data	query posts	100

In part two, the system inputs the queries derived from the test posts, and then segments the Chinese words and removes the stop words. Finally, the model produces the topic words set *T* the given query *q*, being prepared for retrieving the similar post from the dataset.

2.2 Matching

The main work of this stage is to find appropriate post-comment pairs for the given query q. The basic idea here is to find the posts being similar to the query q and use their comments as the candidates [1].

The system firstly retrieves a number of candidate posts by the topic words set *T*. And then, the system evaluates the query-posts similarities, returns the top-*N* most similar posts, and produces the set $P_q^{reduced}$ of reduced candidate posts.

Here we use a simple VSM for measuring the similarity between the query q and the candidate post p in the $P_q^{reduced}$. The score is more close to 1 if the two texts are more similar.

$$sim_{Q2P}(q,p) = \frac{\vec{q}^T\vec{p}}{\|\vec{q}\|\|\vec{p}\|} \quad (1)$$

Where $\vec{q}$ and $\vec{p}$ are the TF-IDF (Term Frequency–Inverse Document Frequency) vector of q and p respectively. The assumption here is that if a post p is similar to the query q, its associated comments might be appropriate for q [1].

Moreover, the system further picks post-comment pairs from $P_q^{reduced}$. Our system measures the similarity between q and each of comments in post-comment pairs, and then produces the candidate comments set *C*.

We also use the VSM for measuring the similarity between the query q and the candidate comment c in *C*.

$$sim_{Q2C}(q,c) = \frac{\vec{q}^T\vec{c}}{\|\vec{q}\|\|\vec{c}\|} \quad (2)$$

Where $\vec{q}$ and $\vec{c}$ are the TF-IDF vector of q and c respectively.

In most cases, a good response has many common words as the query. For example, in Table 2, when the query and the candidate responses both contain "青岛", it is a strong signal that they are about similar topics. This measure, requiring no learning and works on infrequent words, is easy and helpful in finding relevant responses.

Table 2. Post and comments containing common words

Query *q*	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
Comment *c1*	欢迎来到青岛
Comment *c2*	我对青岛印象也很不错
Comment *c3*	你们在青岛？

2.3 Ranking

The system uses a linear ranking function defined in Formula (3) to further evaluate all the comments in *C*, and assigns a ranking score to each candidate comment.

$$rank(q,c) = \alpha * sim_{Q2P}(q,p) + (1-\alpha) * sim_{Q2C}(q,c), 0 \le \alpha \le 1 \quad (3)$$

In Formula (3), α is the adjustable parameter.

Then, the system ranks the candidate comments based on their scores and returns the comments with top-10 comments to the given post.

3. EXPERIMENTAL RESULTS

In this paper, our work focuses on the Chinese subtask that given 100 test posts and then the system is asked to provide a ranked list of ten results for each given post. The comments must be those from the repository [4]. The experiment has been made on the basis of the NTCIR-12 retrieval repository.

The official evaluation measures of the STC task are graded relevance IR evaluation measures for navigational intents and these evaluation measures are computed using the NTCIREVAL tool [4]. All the results of all the runs from teams are pooled to imitate manual annotation by the NTCIREVAL tool. The comments are labeled with L0 (inappropriate), L1 (appropriate in some contexts), and L2 (appropriate) by multiple judges.

We submitted one result of our system for STC task in Chinese and the official evaluation results are listed in the following Table 3.

Table 3. Part of official STC results

Run	WUST-C-R1	BUPTTeam-C-R4
Mean nDCG@1	0.0567	0.3567
Mean nDCG@1	0.1218	0.5082
Mean nERR@10	0.0980	0.4945

Table 3 only lists the official evaluation results of our group and BUPTTeam. There are another 43 official evaluations from 16 teams for Chinese subtask, most of which better than ours. Comparing with the other runs, our system fails to produce the desired result.

We find out that only 38 test posts returned appropriate comments by analysing the offical results of our system. So the average score is very low, and even many posts have zero score. One reason is the assignment for the adjustable parameters. Our model only selects the top-*N* similar posts and produces $P_q^{reduced}$ candidate post-comment pairs. In our experiment, we have assigned *N* to 10. As a result, the post-comment pairs in $P_q^{reduced}$ lack of diversity and miss some relevant posts. Table 3 shows an example from test dataset, and we can find that the query and the candidate posts are relevant but they are ranking after ten in the ranking list. On the other hand, the ajustable parameter α has been firmly assigned to 0.2 and we have not carried out the other tries. The values of *N* and α can be not suitable in our formal run. The detail is shown in Figure 2, where two lines denote the score by nERR@10 measure and the number of appropriate comments searched for the given query respectively.

Table 4. An example of relevant posts ranking

Query *q*	刚刚看到一个骑摩托车的小伙子被撞飞，阳江的交通啊。
candidate posts	**rank**
刚才差一点被狂奔逆行的三轮摩托车撞倒！	17
北京交通，挑战你的膀胱	24
今天有点悲催和幸运，差点被一辆闯红灯的卡车撞飞。	110

There are other reasons to be taken into account when analysing why our system has not worked well. One reason is that our model

uses the simple VSM rather than semantic similarity for measuring query-posts and query-comments similarities. If there are no words overlapping between the candidate comments and the query, the comments will rank very low even they are appropriate. Table 5 shows a real example of the no words overlapping between the query and the comments. Two candidate responses are suitable to the query, while their ranks are very low. The main reason is the no words overlapping between the candidate responses and the query, although there is one common word "瑜伽" between the query and the candidate post.

Table 5. An example of suitable comments ranking

Query *q*	有个小伙子练瑜伽。然后旁边的人都吐血了。
Candidate post	呐呐，坐的太久没动了那就来做个瑜伽吧~~
Comment1	这身子软得
Comment2	他好软啊

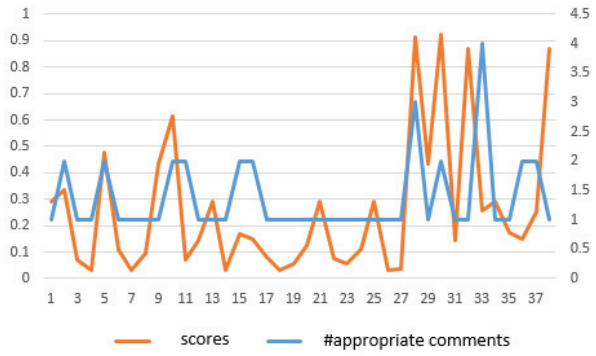

Figure 2. Evaluation scores and the number of appropriate comments for each answered test query

In addition, the high dimensionality and the sparsity are the characteristic of the short text. Some of the meaningful words maybe regarded as the stop words. As a consequence, the segmentation results far away from achieving ideal effect, which weak the system more or less.

After tuning the parameters for many times, the results of the additional evaluation of our system outperform the official ones of our run, which is shown in Table 4. Note that all the results are based on the measures of nERR@10 and the test dataset. The additional evaluation results show that the reasons mentioned above are reasonable and our model can make remarkable progress when the parameters N and α have been assigned to 100 and 0.6 respectively.

However, the best result of our model is very low compared with other groups. Our model is simple while other groups' model combined with other measures. For example, BUPTTeam Group using PageRank combined with similarity, many groups using features extracted or topic words picked through machine learning.

Table 6. Results of our model by parameters tuning

Scores	α = 0.4	α = 0.5	α = 0.6
N = 100	0.198	0.217	0.221
N = 150	0.200	0.209	0.222

4. CONCLUSIONS

In this paper, we have described our model based on VSM for STC task in Chinese. We also analyzed our result submitted and then adjusted parameters which outperformed the former.

We feel that there are two important ways to improve the efficiency of our model for STC task. We need to enhance the accuracy by combining with other models, such as the topic-words, and to consider matching between query and response in terms of semantic relevance, speech act, and entity association.

In the future, besides on the basis of information-retrieval, we also would like to generate the appropriate and human-like response derived from what we searched from the post-comment pair repository.

ACKNOWLEDGMENTS

The work presented in this paper is partially supported by the Major Projects of National Social Foundation of China under Grant No. 11&ZD189 and Natural Science Foundation of China under Grant No. 61402341.

REFERENCES

[1] Z. Ji, Z. Lu, H. Li. 2014. An information retrieval approach to short text conversation. *Eprint Arxiv*

[2] H. Wang, Z. Lu, H. Li, and E. Chen. A dataset for research on short-text conversations. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013:935–945.

[3] M. Efron, J. Lin, J. He, et al. Temporal feedback for tweet search with non-parametric density estimation. *International ACM SIGIR Conference on Research & Development in Information Retrieval.* ACM, 2014:33-42.

[4] L. Shang, T. Sakai, Z. Lu, H. Li, et al. Overview of the NTCIR-12 Short Text Conversation Task. *In Proceedings of the 12th NTCIR Conference.*

[5] M. Richard, M. Craig. Relevance in microblogs: enhancing tweet retrieval using hyperlinked documents. *Conference on Open Research Areas in Information Retrieval.* 2012:189-196.

[6] T. Sakai, L. Shang, Z. Lu, and H. Li. Topic set size design with the evaluation measures for short text conversation. In Proceedings of AIRS 2015 (LNCS 9460), 2015.

[7] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In Proceedings of ACL 2015, 2015:1577–1586.