BJUT at TREC 2016: LiveQA Track

Youjun E, Weitong Chen, Zhen Yang*
College of Computer Science, Faculty of Information, Beijing University of Technology, China
yangzhen@bjut.edu.cn

Abstract

The paper presents the BJUT's liveQA system participating the TREC 2016. The Trec LiveQA track continues to use the last year's instruction, requiring that the system is able to answer the questions which had not been solved in one minutes based on Yahoo! Answers. Our work: (1) The key words are abstracted from the questions, so that more relevant questions will be returned. (2) The system searches in a larger scope on Yahoo! Answers to find the most accurate answers. (3) The answers should be detect whether they are more suitable for answering the given questions. The experiment results are presented at the end of the paper.

Introduction

The automated question answering (QA) track, which has been one of the most popular tracks in TREC for recent years, has focused on the task of automatically answering questions posed by humans in a natural language. The track primarily dealt with factual questions, and the answers provided by participants were extracted from a collection of news articles. While the task evolved to model increasingly realistic information needs, addressing question series, list questions, and even interactive feedback, a major limitation remained: the questions did neither come from real users, nor in real time(Robertson and Walker 1997; Mikolov et al. 2013).

The Trec LiveQA track mainly aims at providing the automatic answers for questions posed by humans in a natural language. There is also an additional demand that extracts the keywords from the question. This track revives and expands the QA track, focusing on live question answering for real-user. Real user questions, extracted from the stream of most recent questions submitted on the Yahoo Answers (YA) site that have not yet been answered by humans, will be sent to the participant systems. The systems will provide an answer in real time. The list of YA categories is limited to a certain range, which includes Arts & Humanities, Beauty & Style, Health, Home & Garden, Pets, Sports and Travel. The question will be provided every minute for a whole day. The returned answers is restricted to 1000 characters and will later be judged by TREC editors on a 5-level Likert scale.

This paper introduces our liveQA system which we use to accomplish the Trec LiveQA track task answering the questions in real time. Since the questions are all from Yahoo Answer, we assume that the questions input into the system have been asked by other people previously, and these similar questions have already had best answers. So we transfer the task from answering the questions to choosing the best answers by similar questions. We don't use any search engine, because we think the answer in Yahoo! Answers is more general.

System Overview

The input of the system is the questions which are written in spoken language. At first, the questions need to be pre-processed. The number of the questions' words should be shorten and the key words should be abstracted. Second, the results after last step should be sent to the related search engine, to get some similar questions and their best answer. In order to get more similar question, the system choose two ways to get more than five similar questions. Then the system formulate the answers and choose the best one as the final answer. In addition, the answer we have chosen need to be expanded. Finally, the keywords need to be selected as the additional function(Banea et al. 2012; Yi, Wang, and Lan 2015; Toba et al. 2014; Shah and Pomerantz 2010).

Question Processing Part

The system receives four parts during evaluation which include qid, title, body and category. Our system doesnt use category because we find that some similar questions are divided into several different categories. We use qid to filtering out from the search results with the same qid because the qid we get doesn't have any answers. Since most questions dont have body and some titles have enough information, the system will only use body when the body has enough meaning words, else add some body words as search requirements.The method is:

```
if(title ≥ T):
    use title as question
else:
    use title and some of the body as question
```

We will delete these words, such as stop words, which are not useful. They don't have any meanings, so they can't take effect on finding best answers. At the same time, some verbs

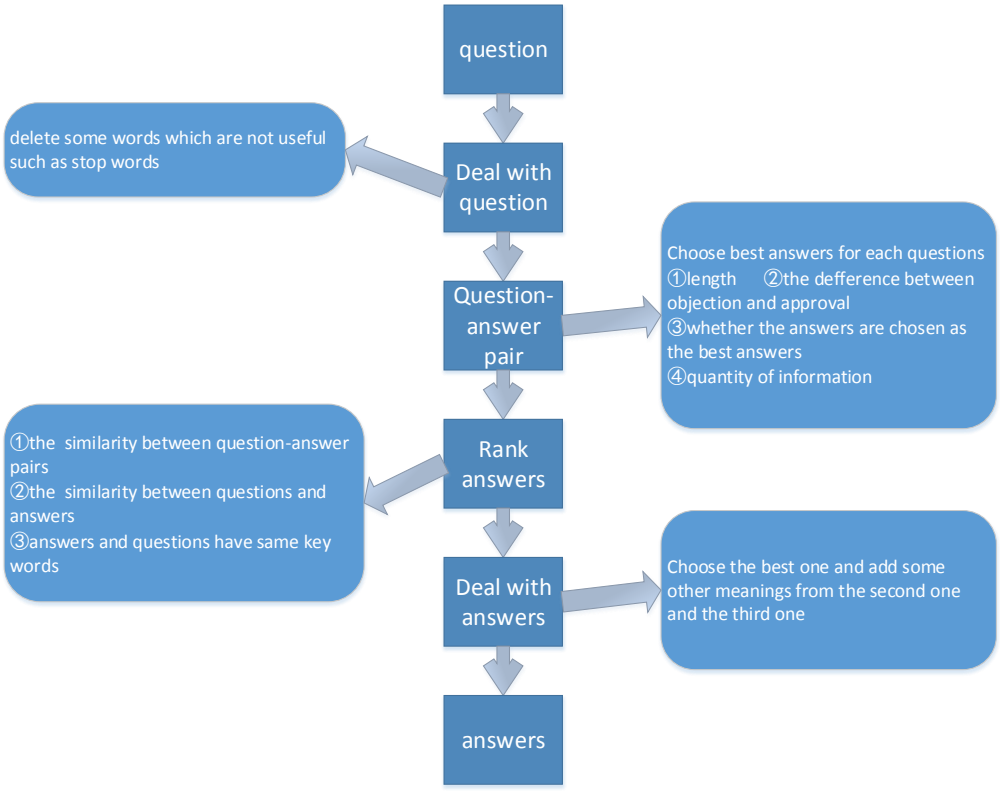

Figure 1: System Framework.

and nouns are taken place with synonyms in order that more similar questions will be found.

Clue Retrieval Part

Yahoo! Answers is chosen as the corpus from which we can find the answers. The system will use the result after the last step as the search requirements to find some similar questions first. Then the system will find all the best answers of these questions. If some questions dont have best answers, the system will formulate all the answers of its question and choose one as the best answer according to the similarity between the question and the answer, the length of the answer and the number of the supporters. If the question dont have similar ones, the system will find its related problems as the similar questions in order that any question given to the system will find its similar questions. As a result, every questions will find about five questions and their best answers which can be used to do next step and be able to get better results.

Answer Processing Part

After the last step, the system gets about five similar questions and their best answers. In this part, the best answer will be chosen from about five answers according the similarity between the input question and their own question, the similarity between the answer and their own question and so on. The system uses cosine similarity to calculate the level of similarity. At the same time, the limitation of the result is 1000 characters, so if the length of the best answer is less than 500 characters, the answer will be replenished according to the answer with the second highest score. The system will choose different meanings in the second and third highest score. We think this step can help the results to get good grade.

Evaluation

In this years LiveQA evaluation, all the questions were scored using 5-level scale:

- -2: non-readable
- 1: poor
- 2: fair
- 3: good
- 4: excellent

The evaluation measures used are:

- avg-score (0-3): average score over all queries (transferring 1-4 level scores to 0-3, hence comparing 1-level score with no-answer score, also considering -2-level score as 0)
- succ@i+: number of questions with i+ score (i=1..4) divided by number of all questions

- prec@i+: number of questions with i+ score (i=2..4) divided by number of answered only questions

Our team hasn't receive any results until now, maybe we didn't return any questions in a right way or other reasons. However, during evaluation, we think our results are good. But we won't lose confidence. We think we can get good results next year.

Conclusion

We presented the BJUTs liveQA system above. This is my first time participation in LiveQA task. However, the results of our system are not satisfactory. On the one hand, we haven't receive any results to judge how the system works during evaluation. On the other hand, without sufficient time, we haven't done a satisfactory model. We will get over them next year, and hope to get good result next year. Because of the lesson this year, we have more experience for the next year evaluation.

The LiveQA track revives the task of automatic question answering in TREC. It provides an opportunity for the participants to try their QA systems on real-world questions, collected from Yahoo! Answers community question answering website. The approach we chose is based on picking key terms from a given question, submitting them to a search engine and extracting an answer.

We would like to improve the procedure of finding answer to a given question by analyzing existing human-generated question-answer pairs. We are hopeful that finding the ways in which an answer is related to the question will help extract more precise answers in the future.

References

Banea, C.; Hassan, S.; Mohler, M.; and Mihalcea, R. 2012. Unt: A supervised synergistic approach to semantic text similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 635–642. Association for Computational Linguistics.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Robertson, S. E., and Walker, S. 1997. On relevance weights with little relevance information. In *ACM SIGIR Forum*, volume 31, 16–24. ACM.

Shah, C., and Pomerantz, J. 2010. Evaluating and predicting answer quality in community qa. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 411–418. ACM.

Toba, H.; Ming, Z.-Y.; Adriani, M.; and Chua, T.-S. 2014. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences* 261:101–115.

Yi, L.; Wang, J.; and Lan, M. 2015. Ecnu: Using multiple sources of cqa-based information for answer selection and yes/no response inference. *In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval* 236.