Finding Semantically Similar Questions Based on Their Answers

Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee
Center for Intelligent Information Retrieval, Computer Science Department
University of Massachusetts, Amherst, MA 01003
[jeon,croft,joonho]@cs.umass.edu

ABSTRACT

A large number of question and answer pairs can be collected from question and answer boards and FAQ pages on the Web. This paper proposes an automatic method of finding the questions that have the same meaning. The method can detect semantically similar questions that have little word overlap because it calculates question-question similarities by using the corresponding answers as well as the questions. We develop two different similarity measures based on language modeling and compare them with the traditional similarity measures. Experimental results show that semantically similar questions pairs can be effectively found with the proposed similarity measures.

Categories and Subject Descriptors

H.3.0 [**Information Search and Retrieval**]: General

General Terms

Algorithms, Measurement, Experimentation

Keywords

Information Retrieval, FAQ retrieval, Language Models

1. INTRODUCTION

Many web sites have question and answer boards or FAQ pages. Retrieval of these human answered questions is very attractive since users can directly obtain answers rather than relevant documents. In such retrieval systems accurate similarity measures between questions are crucial. However, similarity measures developed for documents do not work well for questions because questions are much shorter than documents. Traditional similarity measures for sentences such as the overlap coefficient, Dice coefficient and Jaccard coefficient work poorly when there is little word overlap between sentences.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

Three different types of approaches has been developed in the literature to solve this word mismatch problem as follows: The first approach uses knowledge databases such as machine-readable dictionaries [3]. However, currently there are problems with the quality and structure of knowledge databases. The second approach employs manual annotations or rules, such as AskJeeves[1]. This approach is expensive and hard to expand to other domains. The final approach uses the statistical techniques of information retrieval [1].

We think that the third approach is the most promising if we can have a large number of semantically similar but lexically different question pairs. From these samples, we may extract statistically meaningful patterns to bridge the lexical chasm. The collections of similar question pairs can be further used in many other IR and NLP research areas such as FAQ retrieval, question answering, example based machine translation and so on.

In this paper, we study automatic methods of finding such pairs from existing question and answer collections. Our assumption is if two answers are very similar, then the questions connected to the answers should be semantically similar even though the two questions may be lexically very different. We also study reliable similarity measures between answers.

2. SIMILARITY MEASURES

To find similar answer pairs, reliable similarity measures between answers are required. The lengths of answers vary significantly. Answers can be very short especially for factoid questions. Some answers are very long because sometimes people generating answers just copy multiple related documents from the web. Therefore, any similarity measure seriously affected by length is not appropriate for answers. In this paper, we test three different similarity measures. The first one is the cosine similarity with TF.IDF weights. This measure has been extensively used for various IR and NLP tasks. An advantage of using the cosine similarity is that the measure is symmetric.

The second one is the language modeling technique [2]. The cross entropy between two language models is widely used. However, the cross entropy values are not probabilities. A pair of answers that has a higher cross entropy score than other pairs does not necessarily have stronger semantic connections than the other pairs. For this reason, we do not use cross entropy. Instead, we convert every answer into a

[1]http://www.ask.com

Rank	Cosine	LM-SCORE	LM-HRANK
10	0.00	0.90	0.80
100	0.21	0.67	0.64
1000	0.27	0.41	0.48

Table 1: The ratio of correct answer pairs in top 10, 100 and 1000 positions for each similarity measure.

query and retrieve other answers using the query likelihood language modeling technique. The outputs of the language modeling technique are probabilities and can be used across different pairs of answers. One property of this measure is the scores are not symmetric. Every pair has two different scores depending on which answer becomes a query. In this study, we just pick the maximum value of the two scores. We call this measure LM-SCORE.

The third measure is similar to the second measure in that it uses a language modeling technique. This measure uses ranks instead of scores to resolve the problem of non-symmetric scores. If answer A retrieves answer B at rank r_1 and answer B retrieves answer A at rank r_2, then the similarity between two answers is defined as the reverse of the harmonic mean of r_1 and r_2. $sim(A, B) = \frac{1}{2}(\frac{1}{r_1} + \frac{1}{r_2})$. We call this measure LM-HRANK.

3. EXPERIMENTS

3.1 Environment

We collected 5,200 question-answer pairs from NHN Corp.'s Question and Answer service[2]. All the questions are about email and written in Korean. The average length of questions is 5.9 (words) and the average length of answers is 150.1. There are many semantically equivalent questions in the dataset because many users do not carefully check whether there is the same question in the database before asking their questions. To calculate the cosine similarity and the query likelihood language models, we used the LEMUR[3] toolkit.

3.2 Results

In total, 1,351,700 pairs of answers are possible from 5,200 answers. All of these pairs are ranked according to the three different similarity measures. We manually evaluate the top 1000 pairs for each method. If a question pair connected to an answer pair is semantically identical or very similar, we judge the answer pair is a correct match. Table 1 shows the ratio of the correct matches in the top 10, 100, and 1000 pairs for each similarity measures.

The cosine similarity works poorly because the measure favors short answers. For example, in our dataset, an answer has only two words ("Korean homework") and answer pairs containing this short answer usually have very high cosine similarity scores. Because of this serious problem, the cosine similarity can not be a good similarity measure for answers.

The language modeling technique based measures show good performance. In LM-SCORE, 90% of the answer pairs in the top 10 connect semantically equivalent questions. In the top 100, 67% of the answer pairs are correct matches. LM-HRANK show better results than LM-SCORE in the comparison with the top 1000 pairs. Table 2 shows examples of the question pairs found using the LM-HRANK measure. Each question pair in the examples contains semantically similar questions but questions share very few common terms.

Can I attach a 5 mega byte file in my email? Sending big movie files to my friends over the net by email
Why do we have to use only English for email addresses? Why can't I use Korean in email IDs?
What is the best email service? Who provides the most popular and powerful email accounts?
Who invented email? The first person who used email

Table 2: Examples of question pairs found using the LM-HRANK measure. (English translations).

While LM-SCORE and LM-HRANK show comparable performance, they retrieve different sets of answer pairs. The number of overlapping answer pairs between the top 100 pairs in LM-SCORE and the top 100 pairs in LM-HRANK is only 6. This implies more correct answer pairs can be retrieved when both measures are used together.

We have also tested the use of the scores generated from the traditional TF/IDF model and the Okapi BM25 model as similarity measures. However, the results are not much better than the results of the cosine similarity.

4. CONCLUSIONS

The experimental results show that we can automatically find semantically similar question pairs by measuring similarities between answers. By applying this technique to many different question and answer collections, a large number of similar question pairs can be gathered. We also find language modeling based similarity measures are more appropriate than other similarity measures in calculating similarities between answers. The proposed similarity measures can be used to cluster question-answer pairs and the clusters can be further used to automatically generate FAQs or improve the performances of question and answer retrieval systems.

5. ACKNOWLEDGEMENTS

This work was supported by NHN Corp., the Center for Intelligent Information Retrieval and NSF grant number DUE-0226144. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s)' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, Bridging the lexical chasm: Statistical approaches to answer-finding. *Proceedings of the 23rd annual international ACM SIGIR Conference*, pages 192–199, 2000.

[2] J. M. Ponte, and W. B. Croft, A language modeling approach to information retrieval. *Proceedings of the 21st annual international ACM SIGIR Conference*, pages 275–281, 1998.

[3] K. Hammond, R. Burke, C. Martin and S. Lytinen, FAQ Finder: a Case-Based Approach to Knowledge Navigation. *Proceedings of the 11th Conference on Artificial Intelligence for Applications*, pages 80–86, 1995.

[2] http://www.naver.com

[3] http://www-2.cs.cmu.edu/lemur/