A re-examination of IR techniques in QA system

Yi Chang
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
changyi@software.ict.ac.cn

Hongbo Xu
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
hbxu@software.ict.ac.cn

Shuo Bai
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
bai@ncic.ac.cn

Abstract

The performance of Information Retrieval in the Question Answering system is not satisfactory from our experiences in TREC QA Track. In this article, we take a comparative study to re-examine IR techniques on document retrieval and sentence level retrieval respectively. Our study shows: 1) query reformulation should be a necessary step to achieve a better retrieval performance; 2) The techniques for document retrieval are also effective in sentence level retrieval, and single sentence will be the appropriate retrieval granularity.

1 Introduction

Information Retrieval (IR) and Information Extraction (IE) are generally regarded as the two key techniques to Natural Language Question Answering (QA) System that returns exact answers. IE techniques are incorporated to identify the exact answer while IR techniques are used to narrow the search space that IE will process, that is, the output of the IR is the input of IE.

According to TREC QA overview (Voorhees, 2001; Voorhees, 2002), most current question answering systems rely on document retrieval to provide documents or passages that are likely to contain the answer to a question. Since document-oriented information retrieval techniques are relative mature while IE techniques are still under developing, most of current researches have focused on answer extraction (Moldovan et al., 2002; Soubbotin et al., 2001; Srihari and Li, 1999). There is little detailed investigation into the IR performance which impacts on overall QA system performance. Clarke et al. (2000) proposed a passage retrieval technique based on passage length and term weights. Tellex et al. (2003) make a quantitative evaluation of various passage retrieval algorithms for QA. Monz (2003) compares the effectiveness of some common document retrieval techniques when they were used in QA. Roberts and Gaizauskas (2003) use coverage and answer redundancy to evaluate a variety of passage retrieval approaches with TREC QA questions.

In most of current researches, the granularity for information retrieval in QA is passage or document. What is the potential of IR in QA and what is the most appropriate granularity for retrieval still need to be explored thoroughly.

We have built our QA system based on the cooperation of IE and IR. According to our score and rank on past several TREC conferences, although we are making progress each year, the results are still far from satisfactory. As our recent study shows, IR results in much more loss comparing with IE. Therefore, we re-examine two important questions that have ever been overlooked:

- Whether a question is a good query for retrieval in QA?

- Whether the techniques for document retrieval are effective on sentence level retrieval?

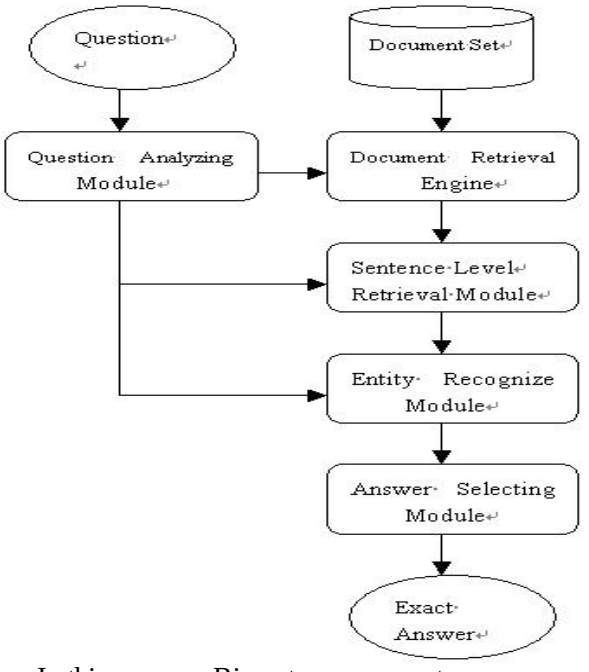
In this paper, we compare some alternative IR techniques and all our experiments are based on TREC 2003 QA AQUAINT corpus. To make a thorough analysis, we focus on those questions with short, fact-based answers, called Factoid questions in TREC QA.

In Section 2, we describe our system architecture and evaluate the performance of each module. Then in section 3, according to the comparison of four document retrieval methods, we find the reason to limit our retrieval performance. We then present in Section 4 the results of four sentence level retrieval methods and in Section 5 we research different retrieval granularities. Finally, Section 6 summarizes the conclusions.

2 System Description

Our system to answer Factoid questions contains five major modules, namely Question Analyzing Module, Document Retrieval Engine, Sentence Level Retrieval Module, Entity Recognizing Module and Answer Selecting Module. Figure 2.1 illustrates the architecture.

Figure 2.1 the Architecture of QA system



In this paper, a Bi-sentence means two consecutive sentences and there is no overlapping between two consecutive Bi-sentences; a phrase means a sequence of keywords or one keyword in a question, where a keyword is a word in the question but not in Stop-word list.

To answer each question, Question Analyzing Module makes use of NLP techniques to identify the right type of information that the question requires. Question Analyzing Module also preprocesses the question and makes a query for further retrieval. Document Retrieval Engine use the question to get relevant documents and selects top n ranked relevant documents. Since the selected documents contain too much information, Sentence Level Retrieval Module matches the question with the selected relevant documents to get relevant Bi-sentences, and selects top m ranked Bi-sentences. Entity Recognizing Module identifies the candidate entities from the selected Bi-sentences, and Answering Selecting Module chooses the answer in a voting method.

In our TREC 2003 runs, we incorporate PRISE and Multilevel retrieval method and we select 50 documents and 20 Bi-sentences. As our recent study shows: among the 383 Factoid questions whose answer is not NIL, there are only 275 questions whose answer could be got from the top 50 relevant documents, while there are only 132 out of 275 questions whose answer could be extracted from the top 20 relevant Bi-sentences. All our statistics are based on the examination of both answer and the corresponding Document ID.

Figure 2.2 Loss Distribution

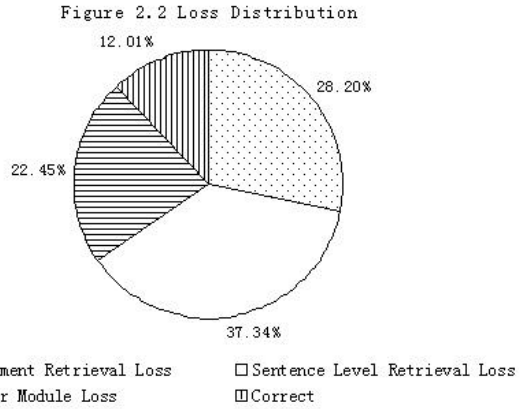


Figure 2.2 illustrates the loss distribution of each module in our QA system. It is amazing that more than sixty percent of loss is caused by IR while we used to take it for granted that IE is the bottleneck of QA system. So we should re-examine the IR techniques in QA system, and we take ac-

count of document retrieval and sentence level retrieval respectively.

3 Document Retrieval Methods

Is a question a good query? If the answer is YES, the retrieval algorithm will determine the performance of document retrieval in QA system. When we implement a document retrieval system with high performance, we could get a satisfactory result. To explore this, we retrieve relevant documents for each question from the AQUAINT document set with four IR systems: PRISE, basic SMART, Enhanced-SMART and Enhanced SMART with pseudo-relevant feedback. In these experiments, we use the question itself as the query for IR system.

3.1 PRISE

The NIST PRISE is a full text retrieval system based on Tfidf weighting, which uses fast and space efficient indexing and searching algorithms proposed by Harman and Candela (1989). To facilitate those participants in the TREC QA Track who do not have ready access to a document retrieval system, NIST also publishes a ranked document list retrieved by PRISE, in which the top 1000 ranked documents per question are provided.

3.2 Basic SMART

SMART (Salton, 1971) is a well-known and effective information retrieval system based on Vector Space Model (VSM), which was invented by Salton in the 1960s. Here we give the experiments with basic SMART and compare its performance with PRISE to make a complete evaluation.

3.3 Enhanced-SMART (E-SMART)

According to the study of Xu and Yang (2002), the classical weighting methods used in basic SMART such as *lnc-ltc* do not behave well in TREC.

In TREC 2002 Web Track, Yang improved the traditional *Lnu-Ltu* weighting method. Following the general hypothesis that relevance of a document in retrieval is irrelevant to its length, Yang thought over the normalization of the query's length. He found that the length of a query should be emphasized much more, but not weakened according to the hypothesis above. When he added the modified *Lnu-Ltu* weighting method into the basic SMART system, the performance of document retrieval was greatly improved. In TREC 2002 Web Topic Distillation Task, this method has been proven to be very effective and efficient.

Inspired by the success in Web Track, we studied the performance of Enhanced-SMART system in TREC QA Track.

3.4 Enhanced-SMART with pseudo-relevant feedback (E-SMART-FB)

We think one of the difficulties of IR in QA system is that the number of keywords is too limited in a question. It is natural to expand the query with pseudo-relevant feedback methods.

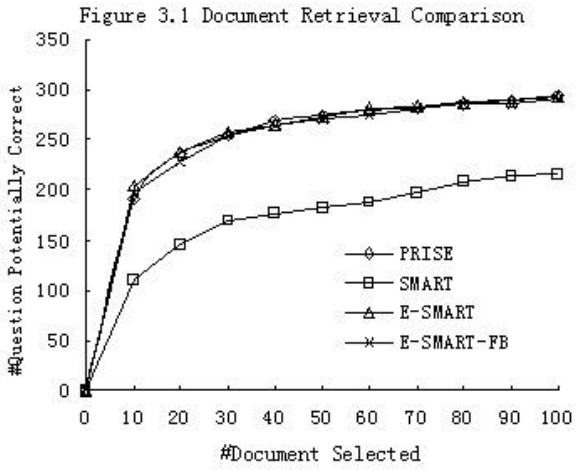
Pseudo-relevant feedback is an effective method to make up for lack of keywords in a query. In our pseudo-relevant feedback, several keywords with the highest term frequency in the top k ranked documents returned by the first retrieval are added into the initial query, and then we make a second retrieval. In the experiments on Web Track after TREC 2002, we correct a bug in SMART feedback component. Our experiments shows that the evaluation performance in Web Topic Distillation Task can outperform that of any other system when we introduce pseudo-relevant feedback into the Enhanced-SMART.

We try to explore the potential of document retrieval in TREC QA Track, so we also do experiments in QA document retrieval by Enhanced-SMART with pseudo-relevant feedback.

3.5 Performance Comparison

Figure 3.1 shows the comparison of four document retrieval methods. X coordinate means the number of documents we select to make statistics. Y coordinate means the number of the questions that could be correctly answered, and here we mean the exact answers of these questions are contained in the selected documents.

According to figure 3.1, the results of Enhanced-SMART retrieval technique, which proved to be one of the best in TREC Web Track, are almost the same as PRISE; basic SMART performs worst, and the performances of other three methods are similar.



In the figure above, selecting 100 documents the curve representing the basic SMART reaches its peak performance of 216, and Enhanced-SMART with pseudo-relevant feedback to 291, Enhanced-SMART to 293. PRISE reaches the best result among them: 294 out of 383 questions could be answered correctly, and the accuracy is 76.76%. Since the document retrieval is still the first step in QA system, such kind of performance is far from satisfactory. Selecting more documents could increase the number of potential correct questions, however, it is likely to impair the accuracy of the following step because the documents which are lower relevant to the question probably contain more noise. That the slope of the curve in the figure is becoming shallower illustrates our analysis.

From TREC 2002 QA overview (Voorhees, 2002), the best system has a final accuracy of more than 70%, which is close to our document retrieval performance. This shows that there should exist a great increase in our document retrieval. It is well known that the retrieval result is based on two factors: the retrieval algorithm and the query. Since the retrieval algorithm has proved to be very effective, while the results it used in TREC QA document retrieval is much lower than expectation, we think the point is the latter: a question is not equal to a good query.

As our further experiments show, the pseudo-relevant feedback doesn't take effect, which means the simple statistics-based query expansion is not good. We should think thoroughly about the query reformulation, and more effective techniques should be studied, in particular, we can make use of semantic information to reformulate the query. We use the required NE type of a question in IE to find the candidate answer, however we overlook it in IR because the words where most of the required NE types are derived from are Stop-words or unimportant words with too high frequency. We think that utilizing the required NE type would be our next attempt to reformulate the query.

Many query expansion methods have been proposed in QA system, but we used to regard them as optional. However, our experiments show that the reformulation of query should be necessary to get a better IR result.

4 Sentence level Retrieval Methods

According to the conclusion of Allan (2003): finding relevant sentences from the document is difficult based on VSM, we used to think that the techniques for document retrieval are not suitable to sentence level retrieval, so we proposed a Multi-level method in order to get a better retrieval result. We compare the Multilevel retrieval with the other three methods: Keyword-match retrieval, TFIDF-based retrieval and Enhanced-SMART-based retrieval. We noticed that there are some questions whose answer should be extracted from more than one sentence, so we take two consecutive sentences as the granularity of retrieval, which we called Bi-sentence. To avoid the repetition of information, we define that there is no overlapping sentence between two consecutive Bi-sentences. In following experiments, we retrieve relevant Bi-sentences from the top 50 documents provided by PRISE.

4.1 Keyword-match Retrieval

The algorithm of Xu (2002) is regarded as one of the basic methods to compute the weight of Bi-sentence:

$$weight\_p = b \times count\_k / (count\_q + count\_p)$$

where $weight\_p$ means the weight of the Bi-sentence, $count\_k$ means the number of matching keywords between the question and the Bi-sentence, $count\_q$ means the number of keywords in the question, $count\_p$ means the number of keywords in the Bi-sentence and b is an experiential parameter. The Bi-sentence with a larger weight has the priority to be retrieved.

4.2 TFIDF-based Retrieval

It is very natural to take the similarity between a question and a Bi-sentence as the weight of the Bi-sentence.

We turn the question and the Bi-sentence into vectors with the Tfidf formula:

$$W(t,\vec{d}) = \frac{tf(t,\vec{d}) \times \log((N+1)/n_t + 0.01)}{\sqrt{\sum_{t \in \vec{d}} [tf(t,\vec{d}) \times \log((N+1)/n_t + 0.01)]^2}}$$

where $\vec{d}$ denotes the vector of the question or the Bi-sentence, $W(t,\vec{d})$ denotes the weight of word t in $\vec{d}$, $tf(t,\vec{d})$ denotes the frequency of word t in the text corresponding to $\vec{d}$, N is the number of all Bi-sentences and n_t is the number of Bi-sentences that contain word t.

The weight of the Bi-sentence is the inner product between the question vector and the Bi-sentence vector. The Bi-sentences that have larger similarity with the question will be retrieved.

4.3 Multilevel Retrieval

Vector Space Model takes a document as a vector with each word an element, and words are independent from each other. It seems that VSM has lost much useful information. Therefore, we want to integrate more information to improve sentence level retrieval. What we think to use first is syntax and semantic information including phase, Chunk, POS and so on, but it is difficult to apply such information in VSM model. So we proposed a Multi-level retrieval method.

Our method is based on two assumptions: 1) Bi-sentences that can match a phrase which is made up of more than one keyword are more relevant than those only can match separate keywords. 2) Bi-sentences that can match a phase of a question in original form are more relevant than those only can match in stemmed form.

We make use of the Chunk, Pos and Stem information to apply a four-level method to select candidate Bi-sentences. At each level, we define two kinds of substrings, Compulsory Phrase and Assistant Keyword. Compulsory Phrase is a phrase set in which each element is obligatory to match a Bi-sentence. Assistant Keyword is a keyword set in which each element is optional to match. Those words not belong to the Compulsory Phrase and Stop-word list are regarded as the elements of the Assistant Keyword. We compute the weight of a Bi-sentence as below:

$$weight\_p = b \times count\_a / (count\_q + count\_p)$$

where $weight\_p$ means the weight of the Bi-sentence, $count\_a$ means the number of matching Assistant Keyword between the question and the Bi-sentence, $count\_q$ means the number of keywords in the question, $count\_p$ means the number of keywords in the Bi-sentence and b is an experiential parameter.

At the first level, we take the last Noun Group and the last verb in the last Verb Group as the Compulsory Phrase. And those phrases with initial capital on each word are also regarded as the Compulsory Phrase. At the second level, we move the verb from the Compulsory Phrase to the Assistant Keyword because the verb is not easy to match and we don't fulfill the verb expansion. At the third level, we only leave those phrases composed of successive initial capital words as the Compulsory Phrase. At the last level, the Compulsory Phrase is empty, and all words belong to Assistant Keyword.

All relevant Bi-sentences are ranked by the following rules: the Bi-sentence selected from the higher level has a higher priority, and in the same level, the Bi-sentence with a larger weight has a higher priority. Furthermore, the first level is based on original matching, while the other three levels are based on stemmed matching.

4.4 Enhanced-SMART-based Retrieval

Whether an effective document retrieval technique is still successful in sentence level retrieval? Since the Enhanced-SMART proves to be an effective document retrieval system in Web Track, we attempt to study its performance in a small granularity.

We first construct Bi-sentences from the top ranked 50 documents retrieved by PRISE. Then we take each Bi-sentence as a document, and use Enhanced-SMART to make index on them. Finally we use the question itself as a query to retrieve the Bi-sentences most relevant to the question.

4.5 Performance Comparison

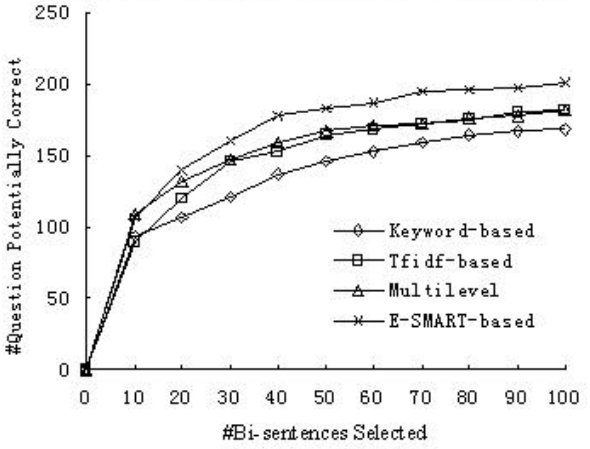
Figure 4.1 shows the comparison of four sentence level retrieval methods in the granularity of Bi-sentence. The performances of Multilevel retrieval

and TFIDF-based retrieval are similar, but what surprises us is that our attempt to Enhanced-SMART-based retrieval makes a hit. It defeats Multilevel retrieval, and gets to a higher performance in sentence level retrieval. Selecting top 100 ranked Bi-sentences, the Enhanced-SMART-based retrieval reaches its peak where answers of 201 questions can be found, while by the Multilevel retrieval, only 181 questions could be correctly answered.

In fact, Enhanced SMART is a variant of Tfidf weighting method based on VSM model. In its *Lnu-Ltu* weighting formula, the length of a document and query is considered well. The experiment results show that Tfidf method based on VSM can also achieve a good performance in sentence level retrieval.

Such a result is beyond our original intention. In figure 4.1, the performance of Multilevel retrieval is better than keyword-match retrieval and simple Tfidf method. It seems that our expectation about Multilevel retrieval is reasonable, but Enhanced SMART overthrows our originalidea.

Figure 4.1 Sentence Level Retrieval Comparison



A reason for us is that we didn't implement a complete Multilevel method, many factors were ignored during the matching at each level, and we also made no use of some more useful information such as NE-type.

We think the limited scale of corpus is another important reason for the unsatisfactory performance of Multilevel retrieval. At the same time, the algorithm in Multilevel method is still too simple, a better scoring strategy is needed to substitute our simple method.

5 The granularities of Retrieval

According to the previous section, we learn that the performance of small granularity retrieval such as Bi-sentence is good using the same techniques as document retrieval. However, we want to make it clear what granularity will perform the best and what is the best performance it can achieve.

In our original opinion, there is no repetition of information in Bi-sentences. However, this method would possibly break two highly relevant sentences into two Bi-sentences and the importance of each Bi-sentence would be weakened by the other sentence in it. Furthermore, if the answer should just be extracted from these two sentences, our endeavor to taking two consecutive sentences into account will be in vain.

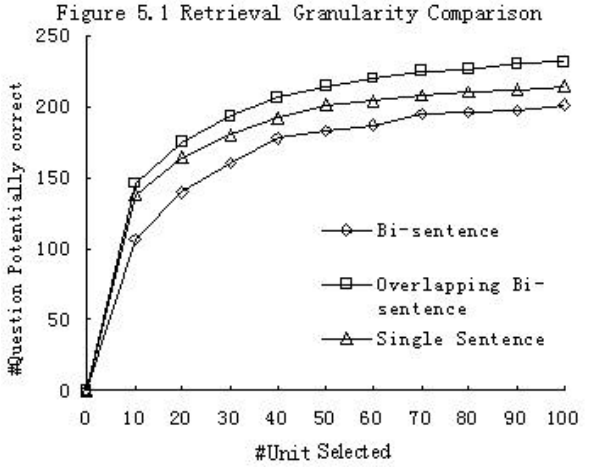
So we examine another retrieval granularity, overlapping Bi-sentence, that is, every two sentences will construct an overlapping Bi-sentence, and every two consecutive overlapping Bi-sentences have one sentence in common. For example:

Four consecutive sentences s_1, s_2, s_3 and s_4 are in a document, and they can construct two Bi-sentences $s_1\_s_2$ and $s_3\_s_4$, while they can build three overlapping Bi-sentences: $s_1\_s_2$, $s_2\_s_3$ and $s_3\_s_4$.

The repetition information is what we try to avoid, so we further consider single sentence as our retrieval granularity. That is, we make index on each single sentence to retrieve in the Enhanced SMART.

Figure 5.1 displays the performance of three granularities of retrieval. Since a Bi-sentence or an overlapping-Bi-sentence contains two sentences, we should compare the number of questions that could be correctly answered between m Bi-sentences and $2m$ single sentences. According to the figure above, Bi-sentence that we used in TREC 2003 QA track performs worst, while the results of the other two granularities are similar: Selecting 50 overlapping Bi-sentences we hit the value of 215 on the represented curve and making a selection of 100 single sentences we get 214; Selecting 100 overlapping Bi-sentences we reach 232 and selecting 200 single sentences we obtain 228.

Figure 5.1 Retrieval Granularity Comparison



Comparing with techniques used in our TREC runs, only 132 out of 275 questions could be correctly answered by selecting 20 Bi-sentences, while utilizing the Enhanced-SMART to retrieve 40 single sentences that are equal in data to 20 Bi-sentences for further process, we could at most correctly answer 192 questions. That is an amazing improvement of 45.45% in sentence level retrieval module.

Selecting 200 single sentences, the accuracy of sentence retrieval reaches 82.91%, that is, 228 out of 275 questions could be correctly answered. Using semantic-based query reformulation, we could probably further improve the performance of sentence retrieval. In a word, what we can affirm is that the techniques for document retrieval are also effective in sentence retrieval.

6 Conclusion

Considering document retrieval and sentence retrieval separately, we achieve an acceptable performance on each module, but the overall performance of IR in QA system is still far from satisfactory because of the cumulated loss in document retrieval and sentence retrieval. Using the 50 top ranked documents produced by PRISE and retrieving 200 single sentences from them, the accuracy of document retrieval is 71.80%, and the accuracy of sentence retrieval is 82.91%, while the overall accuracy of IR is the product of the two module's accuracy, that is only 59.53%.

Since the techniques used in document retrieval and sentence retrieval are the same, we suppose that retrieving sentences from the corpus directly will combine two cumulated loss into one and the overall performance might get better. We will fulfill this idea in the next step.

In conclusion, after re-examining the IR techniques in QA system, we get a satisfactory answer to the two questions presented at the beginning of this article, which will greatly enlighten our future research in QA domain:

- A question is not a good query, and query reformulation should be a necessary step to get a better retrieval performance.
- The techniques for document retrieval are also effective in sentence level retrieval, and single sentence will be the propriety retrieval granularity.

Acknowledge

This research is supported by the National High Technology Research and Development Program (863 program) under contact of 2002AA142110, the Institute Youth Fund under contact of 20026180-24. We give our thanks to all the people who have contributed to this research and development, in particular Xueqi Cheng, Bin Wang, Zhe Yang, Zhifeng Yang, Dongbo Bu, etc.

References

Christof Monz. 2003. *Document Retrieval in the Context of Question Answering*, Proceedings of the 25th European Conference on IR Research.

C.L.A. Clarke, G.V. Cormack, D.I.E. Kisman, T.R. Lynam. 2000. *Question Answering by Passage Selection (MultiText experiments for TREC-9)*, Proceedings of the Text REtrieval Conference 2000.

Dan Moldovan, Sanda Harabagiu, Roxana Girju, Paul Morarescu, Finley Lacatusu, Adrian Novischi, Adriana Badulescu and Orest Bolohan. 2002. *LCC Tools for Question Answering*, Proceedings of Text REtrieval Conference 2002.

Donna Harman, Gerald Candela. 1989. *A Very Fast Prototype Retrieval System Using Statistical Rankin*, SIGIR Forum, 23(3,4): 100-110.

Ellen M. Voorhees. 2001. *Overview of TREC 2001 Question Answering Track*, Proceedings of Text REtrieval Conference 2001.

Ellen M. Voorhees. 2002. *Overview of TREC 2002 Question Answering Track*, Proceedings of Text REtrieval Conference 2002.

Hongbo Xu, Hao Zhang, Shuo Bai. 2002. *ICT Experiments in TREC-11 QA Main Task*, Proceedings of Text REtrieval Conference 2002.

Hongbo Xu, Zhifeng Yang, Bin Wang, Bin Liu, Jun Cheng, Yue Liu, Zhe Yang, Xueqi Cheng, Shuo Bai, 2002. *TREC-11 Experiments at CAS-ICT: Filtering and Web*, Proceedings of Text REtrieval Conference 2002.

I. Roberts, R. Gaizauskas. *Evaluating Passage Retrieval Approaches for Question Answering*, http://www.dcs.shef.ac.uk/research/resmes/papers/CS0306.pdf.

James Allan, Courtney Wade, Alvaro Bolivar. 2003. *Retrieval and Novelty Detection at the Sentence Level*, Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Martin M. Soubbotin, Sergei M. Soubbotin. 2002. *Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach*, Proceedings of Text REtrieval Conference 2002.

Rohini Srihari, Wei Li. 1999. *Information Extraction Supported Question Answer*, Proceedings of the Text REtrieval Conference 1999.

Salton G. 1971. *THE SMART Retrieval System-Experiments in Automatic Document Processing*, Prentice Hall.

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, Gregory Marton. 2003. *Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering*, Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval.

Yi Chang, Hongbo Xu, Shuo Bai. 2003. *TREC 2003 Question Answering Track at CAS-ICT*, Notebook of Text REtrieval Conference 2003.

Yiming Yang, Xin Liu. 1999. *A re-examination of text categorization methods*, Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.