Incorporating Syntactic Information in Question Answering

Xiaoyan Li and W. Bruce Croft

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA
{xiaoyan, croft}@cs.umass.edu

Abstract

Syntactic information potentially plays a much more important role in question answering than it does in information retrieval. The aim of the experiment described in this paper is to study the impact of a particular approach for using syntactic information on question answering effectiveness. The TREC-9 QA track data are used in the evaluation. Our results indicate that a combination of syntactic information with heuristics for ranking potential answers can perform about 10% better than the ranking heuristics on their own.

1. Introduction

Question answering (QA) is a task different from information retrieval (IR) in that it tries to return an exact answer to short fact-based questions instead of a ranked list of documents that are likely to be relevant to users' information needs/queries. Questions submitted to QA systems are full sentences instead of the 2-3 keywords typically given to web search engines. Therefore, syntactic information about how a question is phrased and how sentences in documents are structured potentially provides important clues for the matching of the question and answer candidates in the sentences.

In this paper, we present a particular approach to incorporating syntactic information in question answering. In this approach, both the question and sentences are parsed. The parser used in our system is a statistical parser (SIFT) from BBN [10]. Syntactic information is extracted from the parser output and used in the answer selection process. There are general syntactic clues that apply to all types of questions, such as matching of phrases in the question and the distance between the main verb and an answer candidate in a sentence. There are also some specific syntactic patterns that apply to different types of questions. For example, preferring an answer candidate in a possessive format in a sentence applies to "LOCATION" questions, the questions that require a location as an answer. Adjective noun phrases (NPA) which contain an answer candidate and all query words apply to "PERSON" questions, the questions that require a person name as an answer.

The work presented is this paper is based on the QA techniques and heuristics that are used in a Chinese question answering system--Marsha [7]. Syntactic information is combined with the Marsha heuristics in the new QA system to further improve the accuracy of answer selection. Experiments are done with TREC-9 questions. The experimental results with 60 questions indicate that the combination of heuristics and syntactic information improve the performance of our QA system by 10% compared with our original QA system which used heuristics alone. Currently, in the scoring algorithm for answer selection, the weights of features that are used to calculate the score for each candidate answer are assigned manually.

The rest of the paper is organized in the following way. Section 2 describes answer ranking in QA systems. Section 3 discusses syntactic information that can be used for QA. A particular approach of combining syntactic information with heuristics is given in Section 4. Section 5 provides the experiments and evaluation. Discussions of the experimental results are presented in Section 6. Related work is discussed in Section 7. Finally, conclusions and future work are given in Section 8.

2. Question Answering with Answer Ranking

2.1 Answer Ranking

In question answering, either an answer or a ranked list of answer candidates is expected. In TREC-8 and TREC-9 QA track, a ranked list of up to five (document identifier, answer-strings) pairs for each question is required to be returned. A answer-string is limited to be at most 50 or at most 250 bytes depending on the run type. The interpretation is that answer-string is an answer to the question and doc-id is a document that provides the justification for the answer. Whether an answer or a ranked list of answer-strings is returned, answer-ranking techniques are necessary in QA systems. Typically answer candidates are sorted by their belief scores, which are calculated using heuristics or other techniques. Heuristic ranking techniques are common in QA. The computation of score for an answer window in the LASSO QA system by Moldovan et al. [2] considers heuristics such as the number of matching words in the passage, whether all matching words are in the same sentence, and whether the matching words in the passage have the same order as those in the question. In addition to the above heuristics, the size of the best matching window in a passage and the distance between an answer candidate and the center of the best matching window are considered in our QA system. The best matching window of a passage here is the window that has the most query words in it and has the smallest window size.

2.2 Scoring Algorithm in the Baseline QA System

The baseline QA system consists of three main components: the query processing module, the INQUERY search engine [11], and the answer extraction module. In the query processing module, each question is classified and the type of answer that this question expects is determined. A query is then generated, and is sent to the INQUERY search engine. The search engine takes the query, searches in its data collection and returns the top 10 documents. In the answer extraction module, answer candidates are extracted and their associated scores are calculated. The scoring algorithm is given in Table 1. Four heuristics are considered in the algorithm: the number of matching query words, whether the matching words are in the same sentence, the size of the best matching window and the distance between an answer candidate and the center of the best matching window. The answer candidates then are ranked according to their scores and the answer candidate with the highest score appears at the top of the list.

Table 1: The Scoring Algorithm

```
1. Do the following for each answer candidate in the top 10 passages;
2. Initialize SCORE to 0;
3. Match each query word with words in each passage. Let N stand for the number of matching
   words.  SCORE = SCORE+N;
4. Check whether all matching words in the passage are in a single sentence.
    If yes, then SCORE = SCORE +0.5;
5. Locate the best matching window in the passage and calculate the size of it.
          SCORE = SCORE + N/size of the best matching window;
6. Locate the answer candidate in the passage and calculate the distance between the candidate
   and the center of the matching window in token offset.
      SCORE = SCORE + 0.5/DISTANCE.
```

3. Syntactic Information in Question Answering

The heuristics used in the baseline system make no use of explicit linguistic structure. Syntactic information about how a question is phrased and how sentences in documents are structured potentially provides important clues for the matching of the question and answer candidates in the sentences.

Syntactic information can be extracted from tagging and parsing [9]. Tagging is the task of labeling each word in a sentence with its appropriate part-of-speech like noun, verb, adjective, etc. Parsing is the task of describing the structure of a sentence. The parser output is usually a tree structure with a sentence label as the root, various phrase labels as intermediate nodes, words/symbols in the sentence as leaf nodes and the parent node of a leaf node is the part-of-speech of the word in the leaf node. The parse output of a question can provide potentially more useful information than word-based approaches, where a question is simply viewed as a bag of words or limited features are considered like the order of the words in the question. Syntactic processing extracts information such as part-of-speech tags of words, phrases, and relationships between the words in the question, all of which may be useful information for QA.

In addition, from a parse tree of a sentence, noun phrases, verb phrases, and prepositional phrases etc. are easily recognized. They are usually ignored by general phrase-recognizers that mostly extract proper noun phrases and/or named entities. For example, consider question 294 from Trec9, “Who is the richest person in the world?” Figure 1 represents the parse tree of this question. From this parse tree, the phrases “the richest person” and “in the world” can be extracted. Let’s consider three passages in the documents returned by INQUERY to this question, which are shown in Table 2.

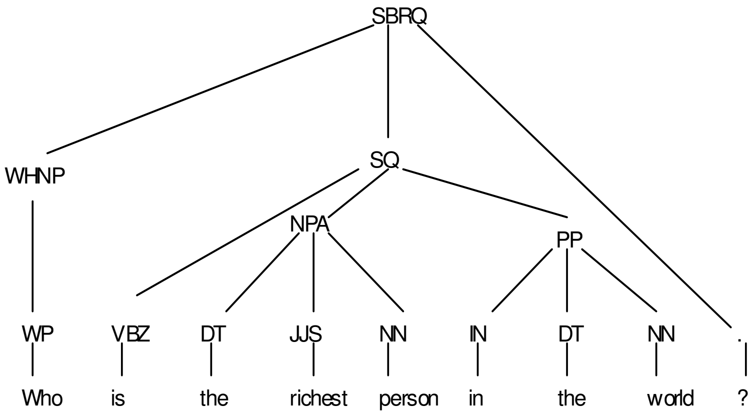

Figure 1: Parsing tree of the question "Who is the richest person in the world?" The actual output from the BBN parser we used is a string that can be easily rebuilt into the tree structure of the question.

If "the richest person" in the question is treated as single query words, then passage 1 and passage 2 may be treated as good passages and "Walton" or "Baker" may be suggested as the best answer to this question although neither of them is the correct answer. With the parse tree of the question, "the richest person" can be extracted and treated as a phrase. Passage 3 will be better than the other two passages when phrase matching is considered and the correct answer "Hassanal" may be extracted.

Table 2: A question and top three passages in the documents returned by INQUERY

Question	*Who is the richest person in the world?*
Passage 1	*Although tops in the U.S., Mr. Walton is the sixth-richest person in the world.*
Passage 2	*Once the richest black person in the world, Baker was destitute shortly before her death. She died in her sleep on the second night of a phenomenally successful comeback show in Paris.*
Passage 3	*As well as being the richest person in the world, Sir Hassanal lives with his relatives in the world's biggest palace _ a complex of buildings built with 38 types of marble on a 300-acre hill near the Brunei River. In case friends decide to stay over, it has 1,778 rooms and 257 toilets.*

The main verb in the question can also be extracted given the parse tree. The relationship between the word "who" and the main verb in the question can be determined. It could be either active or passive. The relationship between an answer candidate and the verb in a sentence and the distance between them are also useful information in the matching of an answer context to a question. For example, for Question 631, “Who won the Nobel Prize in literature in 1988”, the best passage that has the correct answer is as follows:

> *“Afterer Naguib Mahfouz, who won the 1988 Nobel Prize in literature, Abdel-Kuddous was among the best-known novelists in the Arabic language.”*

There are two answer-candidates in this passage: “*Naguib Mahfouz*” and “*Abdel Kuddous*”. “*Naguib Mahouz*” is the correct answer and “*Abdel Kuddousz*” is not. Considering the relationship between the candidate and the main verb “won” and the distance between them, *Naguib Mahouz* can be ranked as the best answer candidate, whereas “*Abdel Kuddous”* is ranked as the top of the list as the best answer candidate to this question in the baseline system which considers only the distance between the candidate and the center of the matching window.

In the proposed QA system, the top 10 sentences are parsed to extract syntactic information. The syntactic information is then combined with heuristics to select more likely answers. While general phrase information and verb related information applies to all types of questions, specific syntactic patterns are also considered for different types of questions. Possessive formats are detected for “LOCATION” questions. Adjective noun phrases are considered for “PERSON” questions. Whether a prepositional phrase with answer candidates modifies the main verb is considered for “LOCATION” and “DATE” questions. All syntactic information is used to adjust the belief score of answer candidates. Section 2.3 describes the details of combining syntactic information with heuristics.

4. Combing Syntactic Information with Heuristic Ranking Techniques

In this section, we will describe in detail how syntactic information is combined with heuristics in our new QA system.

4.1 The Framework of New QA System

Figure 2 presents the relationship between the two systems. The heuristics in the baseline QA system have three main functions. First, they filter out useless passages, which are unlikely to have correct answers. This leaves at most top 10 passages for further parsing and analyzing, thus helping speed up the run time of the new system. Second, answer candidates from the baseline system are potential “back off” answers for the new QA system. Third, the belief score of each answer candidate is a base score that will be adjusted after considering syntactic information.

4.2 Five steps in our new QA system

The new QA system carries out the following five steps, which are given in figure 2:

Step 1: Question Processing.

In this step, the question is parsed using SIFT, a statistical parser from BBN. Adjective noun phrases (NPA), general noun phrases (NP) and prepositional phrases (PP) are extracted from the question. The main verb extracted is a verb in the question but not a stop word. For Who-questions asking for a person, the relationship between the word "who" and the main verb in the question is determined. It could be active or passive depending on whether the person asked is the performer of the action.

Step 2: The baseline QA system is used to find the top 10 passage candidates and their answer candidates.

In this step, an enhancement to the original heuristics is to consider whether the candidate and all the matching words are in the same sentence.

Step 3: Sentence Selection and Parsing.

From the 10 documents returned from INQUERY, one passage is selected from each document using heuristics. Each passage consists of at most 2 sentences. In this step, after considering the number of matching of unique query word and phrases, 10 sentences are selected and sent to the parser.

Step 4: Score Adjusting

In this step, syntactic information from parsing both the question and the sentences is considered and the original belief score of each answer candidate is adjusted accordingly.

Step 5: Answer Ranking

All the answer candidates are ranked by their adjusted belief scores and the top 5 answer candidates are output.

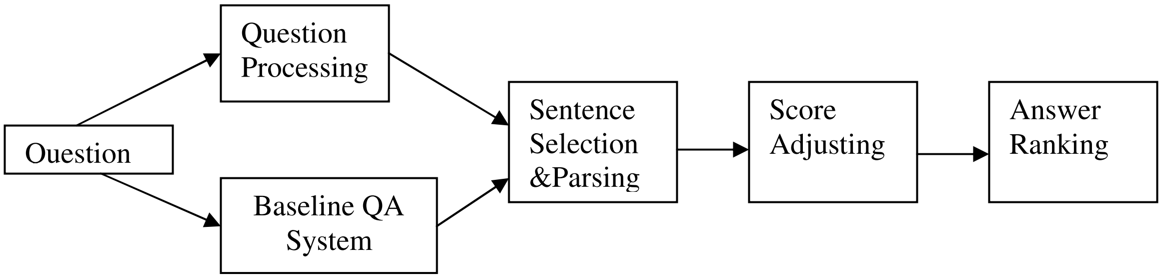

Figure 2: Framework of the New QA System

4.3 Score Adjusting with Syntactic Information

In the step 4 described above, the original belief score of each answer candidate is adjusted. The following factors related to syntactic information are considered and the score is adjusted accordingly, which makes the final belief score for each answer candidate. The weights of each factor considered in this process are currently assigned manually. We are planning to use learning techniques to adjust the weights automatically based on a larger set of question and answer context pairs. The ranking program ranks answer candidates for each question by the belief score and the top 5 are output.

First factor:

Match the sentence with the phrases extracted from the question. If a longer phrase is matched, then the short phrases within it will not be further considered. The belief score is increased by the portion of the sum of the lengths of the matching phrases over the length of the question.

Second Factor:

Consider the distance between the answer candidate and the main verb in token offset. The candidate with the smallest distance is selected as the best answer candidate of the sentence. The belief score increases by 1 over the distance.

Third factor:

Consider the relationship between the answer candidate and the main verb in the sentence for Who-questions. The belief score increases by 0.5, if the relationship is consistent with the relationship in the question.

Fourth factor:

For "PERSON" questions, if all query words and a answer candidate are inside one adjective noun phrase (NPA), the belief score is doubled because we have strong confidence that the answer candidate will be a good answer.

Fifth factor:

For "LOCATION" questions, check the possessive formats such as, "Venezuela's Orinoco" and "Orinoco in Venezuela". If the answer candidate is in such a format, the belief score is then doubled.

Sixth factor:

For "LOCATION" and "DATE" questions, consider whether the candidate is inside a prepositional phrase and modifies the main verb. The belief score is increased by 0.5 + 1/distance. Here the distance is the distance in token offset between the main verb and the answer candidate. The adjustment is done in above way is to simplify the algorithm and accommodate parsing errors.

5. Experiments and Evaluation

The experiments are done with 60 questions from TREC-9. These questions are selected according to the criterion that their question type is "PERSON", "LOCATION", or "DATE" and also whether correct answers can be found in the top 10 documents returned by INQUERY from CIIR. Currently only 60 questions are selected from TREC 9 QA questions because of the limitation of the question classifier and named entity recognizers. The aim of the experiments is to study the impact of a particular approach for using syntactic information on question answering effectiveness. The first experiment we did is running our baseline QA system with these 60 questions. The second experiment is running the new QA system that incorporates syntactic information.

Two evaluation measures are used for comparison. The first evaluation measure is the mean reciprocal answer rank from TREC-9. If the answer is found at multiple ranks, the best rank will be used. If an answer is not found in top five ranks, the score for that question is zero. With this evaluation measure, the QA system incorporating syntactic information achieves 0.849 while the original QA system achieves 0.784. Clearly the new QA system outperforms the baseline. The second evaluation measure is the percentage of correct answers that can be found in the top rank. For 60 questions, there are 40 questions that the correct answer can be found in the top rank using the original QA system. There are 46 questions that the correct answer can be found in the top rank using the new QA system. That indicates the new QA system performs approximately 10% better than the original QA system in terms of this measure.

6. Discussion

There are two conclusions that can be drawn from the experimental results. First, the above experiment indicates that incorporating syntactic information in question answering has a positive impact on question answering effectiveness. Second, it indicates that heuristic ranking provides good back off answers for the new systems. In this sample of 60 questions, the correct answer ranks for 48 questions are unchanged. There are 12 questions where the correct answer ranks are changed, 10 of them are improved and 2 of them are worse. We have studied the cases in which the performance of the system is changed. After examining each question and its corresponding sentences, the main factors that affect the answer selection were discovered. Among the 10 questions where the correct ranks are improved, 3 of them are due to phrase matching, 3 of them are due to the factor that considers whether the candidate and all the matching words are in the same sentence, 2 due to factors that consider the distance and the relationship between the answer candidate and the main verb, and 2 due to possessive formats for “LOCATION” questions. In terms of question types, 4 of them are “PERSON” questions, 4 are “LOCATION” questions and 2 are “DATE” questions. Table 3 gives a summary of the experimental results for different types of questions.

Table 3: Experiment Results

Type of question	Total	Improved	Decreased
Person	29	4	0
Location	20	4	1
Date	11	2	1

The following is an example for questions whose performance is improved.

> Question 249, “Where is the Valley of the Kings?”
> The sentence that has the correct answer is “The newspaper said the remains have not been disturbed since they were sent to the gardens in 1932 by Howard Carter, who discovered the Valley of the Kings at Luxor, Egypt in 1922.” Here “Valley of the Kings at Luxor” is detected by the system as a possessive format. The belief score of “Luxor” is

then doubled. That raises the rank of “Luxor” to the top of all the answer candidates in the new QA system.

The following are the two cases in which the performance decreases:

Question 526, “Where are diamonds mined?” The following are two passage candidates.

Passage 1: “Diamonds are mined in about 20 countries. Australia is the biggest producer in terms of volume or 35 million carats in 1988.”

Passage 2: “After the break-up of the Soviet Union the contract was continued with Rosalmazzoloto, the Russian gold and diamond organization, and an exclusive sales agreement was later signed with Yukutia, the area in eastern Siberia where most Russian diamonds are mined and which is now an autonomous republic in the Russian Federation.”

The heuristic ranking chooses the candidate “Australia” in passage 1 instead of any candidate in passage 2 because Australia is closer to the matching words in the passage. The new system chooses Siberia in passage 2 because it is in the same sentence as the matching words in the passage and it is closer to the matching words than Yukutia. According to the answer contexts in Passage 1 and Passage 2, both Australia and Siberia are correct answers to this question. However, after using the evaluation data provided by TREC-9, only Australia is judged as a correct answer to this question. Siberia is not a correct answer simply because it does not appear in the list of acceptable answers provided by TREC-9. If this misjudgement was corrected, the performance of this question would be unchanged.

Question 851, “When did Mount St. Helens last erupt?” Passage 1 and passage 2 are two passage candidates.

Passage 1: “"Mount St. Helens could erupt again at any time," said Don Swanson, scientist in charge at the USGS observatory in Vancouver Wash. Throughout its recorded history, Mount St. Helens has had active periods that lasted for years with relatively short spans of inactivity. Before the 1980s, the last eruptive period was from 1800 to 1857, with intermittent periods of quiet lasting months or years, according to the USGS western region office in The volcano's most recent eruptions have been quiet, dome-building affairs, in which the mountain pumps out thick lava to increase the size of the crater dome.”

Passage 2: “Mount St. Helens, historically one of the Cascade Range's most active volcanoes, had not erupted since 1857.”

“1980s” is the correct answer to this question. There are 5 query words in the question. Passage 1 has all the 5 query words and Passage 2 only has 4 of them (The query word “last” is not found in the passage). The baseline heuristics choose “1980s” in Passage 1 mainly because of the number of the matching words, while the new system chooses “1857” in the second passage. Two factors here make the belief score of “1857” in the second passage higher than “1980s” in the first passage. One is that the candidate “1857” modifies the main verb “erupt” in the second passage. The other one is that the candidate and the matching words are in the same sentence. Although “1857” is not a correct answer to this question today, it could be correct if the question was asked before 1980s. Actually, the answer to this question is time sensitive and changes when more recent information is available. This issue of time sensitivity will be considered in our future research.

7. Related Work

In this section, we briefly discuss how other researchers have used syntactic information in their QA systems.

Some QA systems do not parse the sentences in documents. For example, Hull [1] used a part-of-speech tagger in his QA system. Basic keywords (e.g. who, where, how etc.) and an associated secondary argument are used to identify question type. The tagger has two functions in this QA system. First, each question is tagged for part of speech and the secondary arguments are extracted using regular expressions defined over sequences of part of speech tags. Second, the function words in the question can be identified by the tagger and then ignored in the process of sentence scoring which scores each sentence according to the number of words it has in common with the question. In Clarke et al.'s [6] QA system, only the question is parsed. The parser here has two functions. One is to generate better queries so that the passage retrieval engine can generate the best candidate passages. The other function is to generate selection rules so that the post processor can select the best 10-byt or 250-byte answers from the passages. The selection rules are patterns for given answer categories (proper, place, time etc.). These patterns generally consist of regular expressions with simple hand-coded extensions.

At the other extreme, some QA systems parse all the text in the corpus, rather than selecting a small subset of sentences that are likely to contain the answer, as is done in our system. Ferret et al.'s [3] QALC system is composed of five parallel modules and a sentences ranking module. The QALC system relies mainly on natural language processing components. Most of the components rely on a tagged version of the corpus by TreeTagger. The patterns of part of speech help assign categories to the questions in the natural language question analysis module, extract terms in the term extraction module and recognize named entities in the named entity recognition module. The parser used by Litkowski[4] is a prototype for a grammar checker. It uses a context-sensitive, augmented transition network grammar of 350 rules. Each sentence in the documents is parsed and databases are constructed by extracting relational triples from the parser output. The triples consist of discourse entities (e.g. numbers, adjective sequences, ordinals, time phrases noun constituents, etc.), semantic relations (roles as agent, theme, location, purpose, etc.), and the governing words, the words in the sentence that the discourse entity stood in relation to. Database triples are also generated for the questions. Matching between the question and sentence database records is done to find candidate sentences, which are more likely to have answers.

Harabgiu et al. [5] makes use of a statistical parser for large real-word text coverage instead of a phrasal parser. The parse trees produced by such a parser can be easily translated into a semantic form. Both the question and the paragraphs returned by the search engine are parsed and transformed into a semantic form. The WordNet semantic net is used to find lexical alternations and semantic alternations. The semantic forms of questions and answers can be unified and thus enable a matching between the conceptual relations expressed in the question and the relations derived from the answer. Our approach differs from this system in that different syntactic patterns are used for specific question types.

In our QA system, we use syntactic information from parsing the questions and sentences to select answer candidates, which are more likely to be correct answers. Heuristics are used to select up to 10 sentences for each question to be parsed. That significantly speeds up the run time. Our QA system focuses on incorporating syntactic information in answer selection.

8. Conclusions and future work

Syntactic information potentially plays a much more important role in question answering than it does in information retrieval. Our experimental results indicate that a combination of syntactic information with heuristics for ranking potential answers can perform about 10% better than the ranking heuristics on their own. The heuristics are also useful for helping filter out useless passages that are unlikely have correct answers, providing "back off" answers and calculating base belief scores that will be adjusted after considering syntactic information.

Currently, in the scoring algorithm for answer selection, the weights of features that are used to calculate a belief score for each candidate are assigned manually. This is mainly due to the fact that we haven't got enough question-answer pairs as a training data set to learn the weights through learning techniques. We are considering incorporating an expanded question classifier and more entity recognizers to classify more questions. Thus a larger set of questions could be used for learning and testing. Then we can choose an appropriate learning technique to learn the weights automatically and do the evaluation on a larger set of questions.

Future work will focus on developing statistic models for question answering which will involve syntactic features. We have already started to develop a statistical model of question answering using the relevance-based model approach [8]. A dynamic aspect of question answering is also worth studying. Question 851 discussed in Section 4 is a good example for this case. For such type of questions, answers may not be decided by one document/paragraph. As new information becomes available, or as new resources are searched, answers may change or be modified. There are other questions that multiple answers are expected.

9. Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD. The grant number is N66001-99-18912. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor.

We also want to express our thanks to people at CIIR for their help, special thanks to David Fisher who implemented the baseline QA system, and Fangfang Feng for his valuable discussions on related research issues.

10. References

[1] D.A. Hull, "Xerox TREC-8 Question Answering Track Report", in proceedings of TREC-8, (1999).

[2] D. Moldovan et al, "LASSO: A Tool for Surfing the Answer Net," in the proceedings of TREC-8, pp 175-183. (1999).

[3] O. Ferret, B. Grau, G. Illouz, C. Jacquwin, and N. Masson, "QALC – the Question-Answering program of the language and Cognition group at LIMSI-CNRS", in proceedings of TREC-8, (1999).

[4] K.C. Litkowski, “Question-Answering Using Semantic Relation Triples”, in proceedings of TREC-8, (1999).

[5] S. Harabagiu, D. Moldovan et al., “FALCON: Boosting knowledge for answer engines”, in proceedings of TREC-8, (1999).

[6] C.L.A. Clarke, G.G. Cormack, D.I.E. Kisman and K. Lynam, “Question Answering by Passage Selection”, in proceedings of TREC-9, (2000).

[7] X. Li and W.B. Croft, “Evaluating Question Answering Techniques in Chinese”, Proceedings of HLT 01, 96-101, (2001).

[8] V. Lavrenko and W.B. Croft, "Relevance-Based Language Models," Proceedings of ACM SIGIR 01, 120-127, (2001).

[9] C. D. Manning and H. Schutze, “Foundations of Statistical natural Language Processing”, The MIT Press, ISBN 0-262-13360-1, 1999.

[10] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, and the Annotation Group, “Algorithms that learn to extract information--bbn: Description of the sift system as used for muc-7”. Proceedings of the Seventh Message Understanding Conference (MUC-7). (1998).

[11] J. P. Callan, W. B. Croft, and S. M. Harding, "The INQUERY Retrieval System", in Proceedings of the 3rd International Conference on Database and Expert Systems. (1992).