The Integration of Lexical Knowledge and External Resources for Question Answering

Hui Yang and Tat-Seng Chua
School of Computing, National University of Singapore
Email: {yangh,chuats}@comp.nus.edu.sg

Abstract

For the short, factoid questions in TREC, the query terms we get from the original questions are either too brief or often do not contain most relevant information in the corpus. It will be very difficult to find the answer (especially exact answer) in a large text document collection because of the gap between the query space and the document space. In order to bridge this gap, there is a need to expand the original queries to include the terms in the document space. In this research, we investigate the integration of both the Web and WordNet in performing local context and lexical correlations to bridge the gap. In order to minimize the noise introduced by the external resources, we explore detailed question classes, fine-grained named entities, and successive constraint relaxation.

1. Introduction

We are participating in this year's Question Answering (QA) main task and it's our first time to take part in TREC. Question answering has recently received attention from many natural language processing communities [1][2][19]. Our goal is to retrieve the exact answers for the short, factoid questions in TREC. In our system, several modules have been developed. They are question processing, external resources adoption, document retrieval, candidate sentence selection and exact answer extraction.

During question parsing, the detailed question classes, answer types, original content query terms and NLP roles of the query terms are analyzed. We derive detailed question class ontology that corresponds to fine-grained named entities. This enables us to extract exact answer from the candidate sentences more accurately.

The original query terms can be used as the basis to locate potential answer candidates in the corpus. However, one major problem of doing this is that the query terms do not have sufficient coverage to locate most answer candidates. This is known as the semantic gap between the query space and document space. In order to bridge this gap, we use the knowledge of both the Web and lexical resources to expand the original query. We first use the original query to search the Web for top N web documents and then extract terms that co-occur frequently in the local context of the N-gram query terms. We next use WordNet to find other terms in the retrieved documents that are lexically related to the expanded query terms. The new query therefore contains terms that are related to the local context in the Web and the lexical context through WordNet. Finally, we use the expanded query to search for answer candidates through the MG system [20].

Candidate answer sentences are selected from the top returned documents and are ranked based on certain criteria to maximize the answer recall and precision. NL analysis is performed on these candidate sentences to extract POS, base Noun Phrases, Named Entities, etc. Answer selection is done by matching the expected answer type to the NL results. The nearest string with the expected answer type in the candidate sentence is returned as the final answer. Figure 1 gives the overview of our system architecture

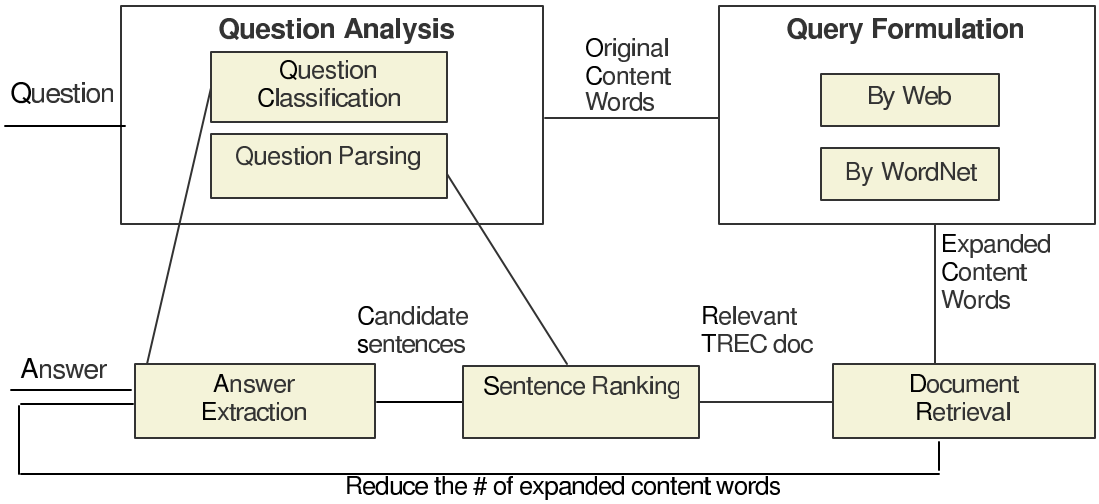

Figure 1: Overview of System Architecture

In this system, we focus on the techniques to expand the original query to locate most answer candidates. The resulting approach is efficient and has been found to be effective. Our experiments on TREC QA main task show good results when combining both local context and lexical information.

2. Question Processing

The purpose of our question processing is to find the specific nature of each question and to make full use of all the information in the question in order to find the best answer.

2.1. Question Classification

Question classification in our system is based on question focus and answer type. A rule-based question classifier is developed to determine the question focus and question class. There are seven main question classes in our system. They are: HUM (Human), LOC (Location), TME (Time), NUM (Number), OBJ (Object), DES (Description) and UNKNOWN (Unknown). The last type UNKNOWN is used to group questions that cannot be categorized into the other classes. Different types of the questions are treated slightly differently in the following answer extraction module.

- Example 1: "Which city is the capital of Canada?" (Q-class: LOC)
- Example 2: "Which province is the capital of Canada in?" (Q-class: LOC)

Obviously, both of the questions belong to the type of *LOC (Location)* and their content words are almost the same, i.e., *capital and Canada*. However, they are expecting different answers, which should fall in different categories, i.e., city or state. The first question's answer will be *Ottawa* but the second's answer should be *Ontario*. In order to detect the subtle differences in the questions, we further classify the first 6 major question classes into 54 sub classes (see table 1 below). Under each main class, there is a special sub-class called *XXX_BASIC*, which is designed for questions that fall in the major class but do not suit any of the sub-classes. Our question classification is similar to the learning classifier developed by Li and Roth [13]. Currently, our rule-based classifier can reach an accuracy of over 98%.

Q-Class	Q-Sub-Class	#Trec11q	#Trec10q	Example
HUM	HUM_PERSON	43	19	Who is the governor of Colorado ?
	HUM_ORG	11	4	What car company invented the Edsel ?
	HUM_BASIC	41	35	Who is Tom Cruise married to ?
LOC	LOC_PLANET	1	2	Which planet did the spacecraft Magellan enable scientists to research extensively ?
	LOC_CITY	18	16	What is the capital city of Algeria ?
	LOC_CONTINENT	3	2	What continent is Scotland in ?
	LOC_COUNTRY	18	3	What country is Berlin in ?
	LOC_COUNTY	3	2	What county is Elmira , NY in ?
	LOC_STATE	3	4	Which state has the longest coastline on the Atlantic Ocean ?
	LOC_PROVINCE	2	2	What province is Calgary located in ?
	LOC_TOWN	2	0	The Hindenburg disaster took place in 1937 in which New Jersey town ?
	LOC_RIVER	3	3	What river is called "China 's Sorrow" ?
	LOC_LAKE	2	2	What is the deepest lake in the world ?
	LOC_MOUNTAIN	1	2	What is the name of the volcano that destroyed the ancient city of Pompeii ?
	LOC_OCEAN	2	1	What body of water does the Colorado River flow into ?
	LOC_ISLAND	3	1	What is the world 's second largest island ?
	LOC_BASIC	50	29	Where is Devil 's Tower ?
NUM	NUM_COUNT	11	12	How many chromosomes does a human zygote have ?
	NUM_PRICE	5	1	How much does it cost to register a car in New Hampshire ?
	NUM_PERCENT	4	8	What percent of the U.S . is African American ?
	NUM_DISTANCE	22	16	What is the height of the tallest redwood ?
	NUM_WEIGHT	0	2	What is the average weight of a Yellow Labrador ?
	NUM_DEGREE	3	7	What is the boiling point of water ?
	NUM_AGE	9	7	How old was Nolan Ryan when he retired ?
	NUM_RANGE	2	0	What is the range for the number of passengers a Boeing 747 airplane can carry ?
	NUM_SPEED	3	5	How fast does a cheetah run ?
	NUM_FREQUENCY	1	0	How often does the United States government conduct an official population census ?
	NUM_SIZE	2	1	What 's the capacity of the Superdome ?
	NUM_AREA	1	1	How much area does the Everglades cover ?
	NUM_BASIC	8	5	How much vitamin C should you take in a day ?
TME	TME_YEAR	24	15	What year was Alaska purchased ?
	TME_MONTH	2	0	In what month are the most babies born ?
	TME_DAY	8	9	What day did Neil Armstrong land on the moon ?
	TME_BASIC	65	23	When was the telegraph invented ?
OBJ	OBJ_CURRENCY	2	5	What is the currency used in China ?
	OBJ_MUSIC	8	2	What was Aaron Copland 's most famous piece of music ?
	OBJ_ANIMAL	5	9	What is the state bird of Alaska ?
	OBJ_PLANT	2	5	What is the major crop grown in Arizona ?
	OBJ_BREED	1	1	What breed was Roy Rogers ' horse Trigger ?

	OBJ_COLOR	2	9	What are the colors of the Italian flag ?
	OBJ_RELIGION	1	1	What is the chief religion for Peru ?
	OBJ_WAR	1	1	What war is connected with the book " Charge of the Light Brigade" ?
	OBJ_LANGUAGE	1	2	What language do they speak in New Caledonia ?
	OBJ_WORK	3	1	Which long Lewis Carroll poem was turned into a musical on the London stage ?
	OBJ_PROFESSION	3	2	What was William Shakespeare 's occupation before he began to write plays ?
	OBJ_ENTERTAIN	2	1	What TV series did Pierce Brosnan play in ?
	OBJ_GAME	2	1	What card game uses only 48 cards ?
	OBJ_BASIC	61	190	What is the chemical formula for sulphur dioxide ?
DES	DES_ABB	9	7	What does CPR stand for ?
	DES_MEANING	1	6	What does " E Pluribus Unum " mean ?
	DES_MANNER	5	4	How did Mahatma Gandhi die ?
	DES_REASON	1	4	What are hiccups caused by ?
	DES_BASIC	14	10	What do you call a baby sloth ?

Table 1 : Question Classes

2.2. Question Parsing

Besides question classes, some important information are also extracted when we parse the question. They are crucial for the later processes. Detailed analysis is performed here in order to get as much useful information as possible. There are several kinds of word groups we extract from the original question. They are:

(1) ***Content Words:*** These include nouns, adjectives, numbers, and some non-trivial verbs, which appear in the question string. Part of Speech tagging is performed before we select the content words. For example: "*What mythical Scottish town appears for one day every 100 years?*" , the content word vector will be $\underline{q}^{(0)}$ **: (mythical, Scottish, town, appears, one, day, 100, years)**

(2) ***Basic Noun Phrases:*** we use noun phrase recognizer to identity all basic noun phrases appear in the question. For the above example, the noun phrase vector **n : (" mythical Scottish town")**

(3) ***Head of the First Noun Phrase:*** It refers to the noun follows the question header (e.g. what, which, how, etc) and carries the main meaning of the question focus. It can be the next noun or the last word in the next noun phrase after the question header. For the above example **h : (town)**. Usually, there is only one such head for each question sentence.

(4) ***Quotation Words:*** For some of the questions, quotations appear in the question string. They should be given special treatment. The string inside quotation marks usually is longer than a noun phrase, sometimes it could be a full sentence. For example, *" What Broadway musical is the song " The Story is Me " from ?"* The quotation word vector will be **u : ("The Story is Me")**

3. Query Expansion

After question processing, we need to locate the relevant documents and sentences from the TREC corpus. The most common way is to apply information retrieval techniques to find the relevant documents and candidate answer sentences. For the short, factual questions in TREC, the query terms we get from the original questions are either too brief or do not fully cover the terms used in the corpus.

Given a short query, $\underline{q}^{(0)} = [q_1^{(0)} q_2^{(0)} .. q_k^{(0)}]$ usually with k<=4, the problem for retrieving all the documents relevant to $\underline{q}^{(0)}$ is that *the query does not contain most of the terms used in the document space to represent the same concept*. Thus there is thus a need to expand the original query to bridge the gap between the query space and document space.

We use general open resources to overcome this problem. The external general resources that can be readily used include the Web, WordNet, Knowledge bases, and Query Logs. Many groups working on QA have recently used the Web [3][4][5][6][8][9][12][18] and WordNet [7][10][11][16][17] as resources for question answering. In our system, we integrate the external resources to expand the query. The new query is then used to look for the relevant documents and sentences in the QA Text Collection.

3.1. Using Web as the Generalized External Resource

The Web is the most rapidly growing and complete knowledge resource in the world now. The terms in the relevant documents retrieved from the Web are likely to be similar or even the same as those in the QA Text Collection since they are both news articles.

Original content words in the question are passed to the online search engine, e.g. Google, to search for documents in the Web. The terms in the relevant web documents that are highly correlated with the original query terms will be considered as candidates to expand the context of the original query. The steps are:

a) Get original query $\underline{q}^{(0)} = (q_1^{(0)}, q_2^{(0)}, \ldots, q_k^{(0)})$.

b) For $\underline{q}^{(0)}$, retrieve the top N documents from the Web.

c) $\forall q_i^{(0)} \in \underline{q}^{(0)}$, extract $\underline{w}_i$, which contains non-trivial words in the same sentence or within *p* words away from $q_i^{(0)}$ in the retrieved web documents.

d) Rank all $w_{ik} \in \underline{w}_i$ by computing its probability of co-occurrence with $q_i^{(0)}$ as:

$$pr(w_{ik}) = \frac{d_s(w_{ik} \wedge q_i^{(0)})}{d_s(w_{ik} \vee q_i^{(0)})} \quad (1)$$

where, $d_s(w_{ik} \wedge q_i^{(0)})$ is the number of instances that w_{ik} and $q_i^{(0)}$ appear together, and $d_s(w_{ik} \vee q_i^{(0)})$ is the number of instances that either w_{ik} or $q_i^{(0)}$ appears.

e) Merge all $\underline{w}_i$ to form $\underline{C}_q$ for $\underline{q}^{(0)}$. Therefore, $\underline{C}_q$ contains the list of words that are highly correlated with the original query from web documents.

3.2. Use WordNet as the Generalized External Resource

The Web can only provide us the words that occur frequently with the original query terms in the local context. It however, lacks information on lexical relationships between these terms. To overcome this problem, we look up WordNet to find words that are lexically related to the original query terms. The glosses, synonyms, and hypernyms are considered to be useful in relating words. In this work, we consider glosses and synonyms only to relate terms. For example, from the glosses,

- Definition of *plant*: A living organism lacking the power of locomotion
- Definition of *animal*: A living organism characterized by voluntary movement

The common concept here is ***living organism***, which will link concept ***plant*** to concept ***animal***.

From WordNet, we can find gloss words $\underline{G}_q$ and synset words $\underline{S}_q$ for $\underline{q}^{(0)}$. If we expand the query by appending all the terms in the glosses and synsets, it tends to be too general and contain too many terms out of context. In general, we need to restrict $\underline{G}_q$ and $\underline{S}_q$ to those terms found in the web documents, i.e., those found in $\underline{C}_q$. Thus we circumvent this problem by using gloss and synset relations to increase the weights of appropriate context terms $w_k \in \underline{C}_q$ by:

- if $w_k \in \underline{G}_q$, increase w_k by α
- if $w_k \in \underline{S}_q$, increase w_k by β, $(0<\beta<\alpha<1)$ *(2)*

The final weight for each term in $\underline{C}_q$ is normalized for ranking. The new query is formed as

$$\underline{q}^{(1)} = \underline{q}^{(0)} + \{\text{top } m \text{ terms from } \underline{C}_q \text{ whose weights are below the selection threshold }\} \quad (3)$$

Currently, we plan to use the Semantic Perceptron Net approach [14] to derive semantic groups in $\underline{C}_q$, $\underline{G}_q$ and $\underline{S}_q$ in order to derive a structured approach to utilize external knowledge.

4. Document & Candidate Answer Sentence Retrieval

We use the MG tool [20] in our system to index the documents. We choose Boolean retrieval because of the short queries and the need to maximize precision. After performing Boolean retrieval by using $\underline{q}^{(1)}$ to retrieve the top M documents (M = 50), if $\underline{q}^{(1)}$ does not return sufficient number of relevant documents, we reduce the extra terms added and repeat the Boolean search. Therefore, we successively relax the constraints to ensure precision in document retrieval.

The sentence is chosen as the basic unit for processing in our system. After performing sentence boundary detection, we use the following criteria to rank the relevance of a sentence to the question: *(Recall from query processing, we extracted* $\underline{q}^{(0)}$, $\underline{n}$, $\underline{h}$, $\underline{u}$*)*. For each Sentence $Sent_j$, we match it with

- quotation words: W_{uj} = % of term overlap between $\underline{u}$ and $Sent_j$
- noun phrases: W_{nj} = % of phrase overlap between $\underline{n}$ and $Sent_j$
- head of first noun phrase: W_{hj} = 1 if there is a match and 0 otherwise
- original content words: W_{cj} = % of term overlap between $\underline{q}^{(0)}$ and $Sent_j$
- expanded content words: W_{ej} = % of term overlap between $\underline{q}^{(1-0)}$ and $Sent_j$, where $q^{(1-0)} = \underline{q}^{(1)} - \underline{q}^{(0)}$

The final score for the sentence is $S_j = \sum a_i W_{ij}$, where $\sum a_i = 1$, $W_{ij} \in \{ W_{uj}, W_{nj}, W_{hj}, W_{cj}, W_{ej} \}$. The top *K* sentences are then selected as the candidate answer sentences based on S_j.

5. Answer Extraction

Finally we perform the tagging of fine-grained named entities [15] on the top *K* sentences extracted from the previous steps. From these sentences, we extract the string that matches the Question classes (answer target) as the answer. Once an answer is found within the top i^{th} sentence, the system will terminate the search for the rest of *(K-i)* sentences. When there is more

than one matching strings in a single sentence, we will choose the string that is nearest to the original query terms. For example: for question "*Where did Dr. King give his speech in Washington?*" , we get:

- Q-class: ***LOC_BASIC***
- <*LOC_BASIC* WASHINGTON> KING-DREAM _ <*LOC_BASIC* WASHINGTON> _ In the <*NUM_PERIOD 35 years*> since Dr . <*HUM_PERSON Martin Luther* King> Jr . delivered his `` I Have a Dream " *speech* at the <*LOC_BASIC Lincoln Memorial*> , how have economic and social conditions changed for <*LOC_CONTINENT African*> Americans ?

For question class ***LOC_BASIC***, we look for all the sub categories under LOC and we will get *WASHINGTON, WASHINGTON, Lincoln Memorial* and *African* as answer candidates. Among them, ***Lincoln Memorial*** is the nearest string to original content word *speech*, and hence is picked as the exact answer.

For some questions, we cannot find any answer. Our solution is to reduce the number of the expanded query terms and repeat the document/sentence retrieval and answer extraction process for up to m iterations (m=5). If we still cannot find an exact answer, NIL is returned as the answer. We call this method *successive constraint relaxation*, which helps to increase the recall while preserving precision.

6. Result Analysis

We answered 290 questions correctly with un-interpolated average precision of 0.61. Figure 2 shows that our system works well for most of the easy questions (right side of the figure), and has reasonable performance for the difficult ones.

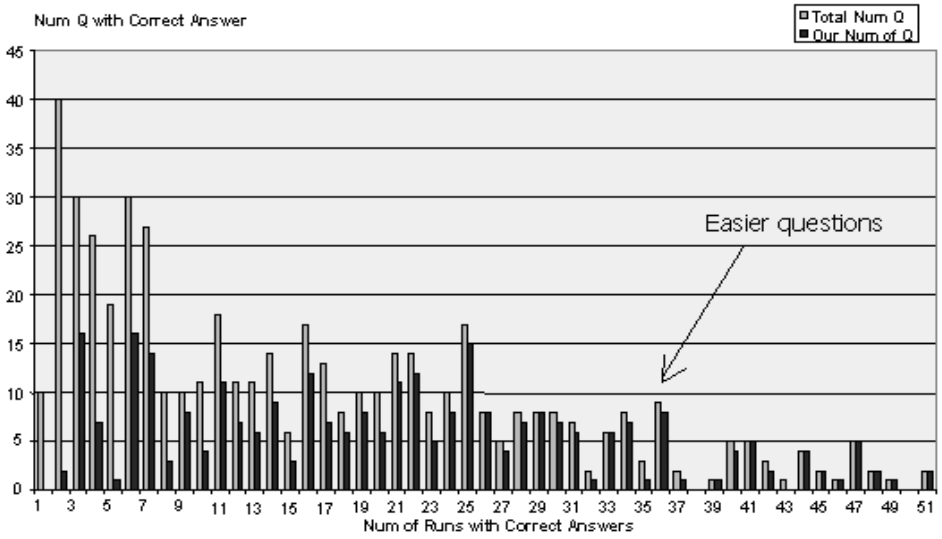

Figure 2: Question Difficulty Distribution

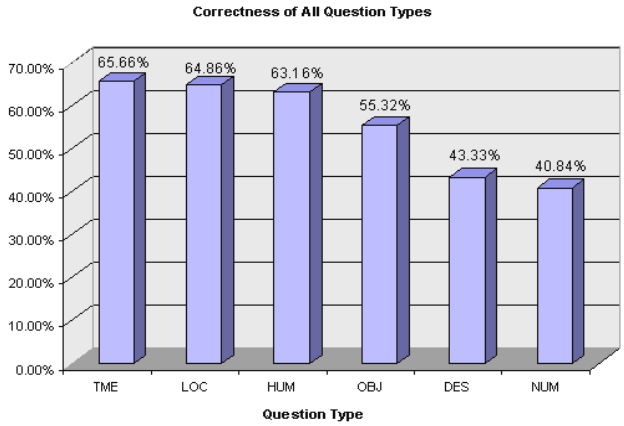

Figure 3: Answer Accuracy of All the Question Types

We also found that the accuracy of the exact answers differ for different type of questions (see Figure 3). For some question classes, like Time, Location and Human, our system gives quite high performance. For Description, Number and Object questions, we still need to find better techniques to improve the performance.

Another problem is that we have too many questions with NIL answers. The precision for recognizing NIL answer is low: 41 / 170 = 0.241, although the recall for NIL answer is satisfactory: 41 / 46 = 0.891. As a result, the overall system recall

(consider both questions with non-NIL and those with NIL answer) is not satisfactory as compared to precision. This is because we use the boolean search to look for relevant TREC documents. Only the documents containing all the query terms are returned. This restriction might be too strict.

7. Future Work

We are currently refining our approach in several directions. First, we are refining our terms correlation by considering a combination of local context, global context and lexical correlations. Second, we are working towards a template-based approach on answer selection that incorporates some of the current ideas on question profiling and answer proofing, etc. Third, we will explore the structured use of external knowledge using the *semantic perceptron net* approach [14]. Our longer-term research plan includes Interactive QA, and the handling of more difficult analysis and opinion question types.

References

[1] AAAI Spring Symposium Series (2002). Mining Answers from Text and Knowledge Bases.

[2] ACL-EACL (2002). Workshop on Open-domain Question Answering.

[3] E. Agichtein, S. Lawrence and L. Gravano (2001). " Learning search engine specific query transformations for question answering" . In Proceedings of the 10th World Wide Web Conference (WWW10), 169-178.

[4] E.Brill,J.Lin,M.Banko,S.Dumais,and A.Ng, " Data-intensive question answering" , Text REtrieval Conference 2001

[5] E. Brill, S. Dumais and M. Banko (2002). "An analysis of the AskMSR question-answering system." In Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).

[6] S. Buchholz (2002). " Using grammatical relations, answer frequencies and the World Wide Web for TREC question Answering" . In Proceedings of the Tenth Text Retrieval Conference (TREC 2001).

[7] J. Chen, A. R. Diekema, M.D. Taffet, N. McCracken, N. E. Ozgencil, O. Yilmazel, E. D. Liddy (2002). " Question answering: CNLP at the TREC-10 question answering track" . In Proceedings of the Tenth Text Retrieval Conference (TREC 2001).

[8] C. Clarke, G. Cormack and T. Lynam (2002). "Web reinforced question answering." In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).

[9] C. Clarke, G. Cormack and T. Lynam (2001). " Exploiting redundancy in question answering" . In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2001), 358-365.

[10] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus and P. Morarescu (2001). " FALCON: Boosting knowledge for question answering" . In Proceedings of the Ninth Text Retrieval Conference (TREC-9), 479-488.

[11] E. Hovy, U. Hermjakob and C. Lin (2002). "The use of external knowledge in factoid QA." In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).

[12] C. Kwok, O. Etzioni and D. Weld (2001). " Scaling question answering to the Web." In Proceedings of the 10th World Wide Web Conference (WWW'10), 150--161.

[13] Xin Li and Dan Roth, " Learning Question Classifiers", In Proceedings of the 19th International Conference on Computational Linguistics, 2002

[14] J. Liu and T. S. Chua, " Building semantic perceptron net for topic spotting" , In Proceedings of 37th Meeting of Association of Computational Linguistics (ACL 2001), Toulouse, France, Jul 2001, pages 370-377

[15] Gideon S. Mann, "Fine-Grained Proper Noun Ontologies for Question Answering", SemaNet'02: Building and Using Semantic Networks, 2002

[16] M. A. Pasca and S. M. Harabagiu (2001). "High performance question/answering" . In Proceedings of the 24th AnnualInternational ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2001), 366-374.

[17] J. Prager, E. Brown, A. Coden and D. Radev (2000). " Question answering by predictive annotation" . In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2000), 184-191.

[18] D. R. Radev, H. Qi, Z. Zheng, S. Blair-Goldensohn, Z. Zhang, W. Fan and J. Prager (2001). "Mining the web for answers to natural language questions" . In Proceeding of the 2001 ACM CIKM: Tenth International Conference on Information and Knowledge Management, 143-150

[19] E.M.Voorhees. " Overview of the TREC 2001 Question Answering Track." In Proceedings of the Tenth Text REtrieval Conference (TREC 2001)

[20] I. Witten, A. Moffat, and T. Bell, "Managing Gigabytes" , Morgan Kaufmann, 1999.