.

The Effectiveness of Dictionary and Web-Based Answer Reranking

Chin-Yew Lin
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
cyl@isi.edu

Abstract

We describe an in-depth study of using a dictionary (WordNet) and web search engines (Altavista, MSN, and Google) to boost the performance of an automated question answering system, Webclopedia, in answering definition questions. The results indicate applying dictionary and web-based answer reranking together increase the performance of Webclopedia on a set of 102 TREC-10 definition questions by 25% in mean reciprocal rank score and 14% in finding answers in the top 5.

1 Introduction

In an attempt to further progress in information retrieval research, the Text REtrieval Conference (TREC) sponsored by the National Institute of Standards and Technology (NIST) started a series of large-scale evaluations of domain independent automated question answering systems in TREC-8 (Voorhees 2000) and continued in TREC-9 and TREC-10. NTCIR (NII-NACSIS Test Collection for IR Systems, TREC's counterpart in Japan) initiated its question answering evaluation effort, Question Answering Challenge (QAC) in 2001 (Fukumoto et al. 2001). Research systems participating in TRECs and the coming QAC focused on the problem of answering closed-class questions that have short fact-based answers ("factoids") from a large collection of text.

These systems bear a similar structure:

(1) Question analysis – identify question keywords to be submitted to search engines (local or web), recognize question types, and suggest expected answer types. Although most systems rely on a taxonomy of expected answer types, the number of nodes in the taxonomy varies widely from single digits to a few thousands. For example, Abney et al. (2000) used 5; Ittycheriah et al. (2001), 31; Hovy et al. (2001), 140; Harabagiu et al. (2001), 8,797. These taxonomies were mostly based on named entities and WordNet (Fellbaum 1998). Special types such *definition* questions (ex: "*What is an atom?*") were added as necessary.

(2) Passage or Sentence retrieval – this aims to provide a text pool of manageable size for extracting candidate answers. Most top performing systems in TRECs use their own retrieval methods for passages (Brill et al. 2001; Clarke et al. 2001; Harabagiu et al. 2001) or sentences (Hovy et al. 2001).

(3) Candidate answer extraction – extract candidate answers according to answer types. If the expected answer types are typical named entities, information extraction engines (Bikel et al. 1999, Srihari and Li 2000) are used to extract candidate answers. Otherwise special answer patterns are used to pinpoint answers. For example, Soubbotin and Soubbotin (2001) create a set of 6 answer patterns for definition questions.

(4) Answer ranking – assign scores to candidate answers according to their frequency in top ranked passages (Abney et al. 2000; Clarke et al. 2001), similarity to candidate answers extracted from external sources such as the web (Brill et al. 2001; Buchholz 2001) or WordNet (Harabagiu et al. 2001; Hovy et al. 2001), density, distance, or order of question keywords around the candidates, similarity between the dependency structures of questions and candidate answers (Harabagiu et al. 2001; Hovy et al. 2001; Ittycheriah et al. 2001), and match of expected answer types.

In this paper, we describe an in-depth study of answer *reranking* for definition questions. Definition questions account for over 100 (20%) test questions in TREC-10. They are not named entities that have been the cornerstones of many

.

high performance QA systems (Srihari and Li 2000; Harabagiu et al. 2001).

By reranking we mean the following. Assume a QA system such as Webclopedia (Section 3) provides an initial set of ranked candidate answers from the TREC corpus. The ranking is based on the IR engine's passage or sentence match scores. One can then measure the effectiveness of utilizing resources such as WordNet or the web to rerank the initial results, hoping to achieve better mean reciprocal rank (MRR) and percent of correctness in the top 5 (PTC5).

Answer reranking is often overlooked. The answer candidates (<= 400 instances per question) generated by Webclopedia from TREC corpus included answers for 83% of 102 definition questions used in this study (the TREC-10 definition questions). However, Webclopedia ranked only 64% of them in the top 5, giving an MRR score of 45%. If a perfect answer reranking function had been used, the best achievable MRR would have been 83% (an 84% increase over the original 45%).

Section 2 gives a brief overview of TREC-10. Section 3 outlines the Webclopedia system. Section 4 defines definition questions and describes our dictionary and web-based reranking methods. Section 5 presents experiments and results. We conclude with lessons learned and future work.

2 TREC-10 Q&A Track

The main task of the TREC-10 (Voorhees and Harman 2002) QA track required participants to return a ranked list of five answers of no more than 50 bytes long per question that were supported by the TREC-10 QA text collection. The TREC-10 QA document collection consists of newspaper and newswire articles on TREC disks 1 to 5. It contains about 3 GB of texts. Test questions were drawn from filtered MSNSearch and AskJeeves logs. NIST assessors then sifted 500 questions from the filtered logs as test set. The questions were closed-class fact-based ("factoid") questions such as "*How far is it from Denver to Aspen*?" and "*What is an atom?*". Mean reciprocal rank (MRR) was used as the indicator of system performance. Each question receives a score as the reciprocal of the rank of the first correct answer in the 5 submitted responses. No score is given if none of the 5 responses contain a correct answer. MRR is then computed for a system by taking the mean of the reciprocal ranks of all questions.

Besides MRR score, we are also interested in learning how well a system places a correct answer within the five responses regardless of its rank. We called this *percent of correctness in the top 5* (PCT5). PCT5 is a precision related metric and indicates the upper bound that a system can achieve if it always places the correct answer as its first response.

3 Webclopedia: An Automated Question Answering System

Webclopedia's architecture follows the principle outlined in Section 1. We briefly describe each stage in the following. Please refer to (Hovy et al. 2002) for more detail.

(1) Question Analysis: We used an in-house parser, CONTEX (Hermjakob 2001), to parse and analyze questions and relied on BBN's IdentiFinder (Bikel et al., 1999) to provide basic named entity extraction capability.

(2) Document Retrieval/Sentence Ranking: The IR engine MG (Witten et al. 1994) was used to return at least 500 documents using Boolean queries generated from the query formation stage. However, fewer than 500 documents may be returned when very specific queries are given. To decrease the amount of text to be processed, the documents were broken into sentences. Each sentence was scored using a formula that rewards word and phrase overlap with the question and expanded query words. The ranked sentences were then filtered by expected answer types (ex: dates, metrics, and countries) and fed to the answer extraction module.

(3) Candidate Answer Extraction: We again used CONTEX to parse each of the top *N* sentences, marked candidate answers by named entities and special answer patterns such as definition patterns, and then started the ranking process.

(4) Answer Ranking: For each candidate answer several steps of matching were performed. The matching process considered question keyword overlaps, expected answer types, answer patterns, semantic type, and the correspondence

.

of question and answer parse trees. Scores were given according to the goodness of the matching. The candidate answers' scores were compared and ranked.

(5) Answer Reranking, Duplication Removal, and Answer output: For some special question type such as definition questions (e.g., "*What is cryogenics?*"), we used WordNet glosses or web search results to rerank the answers. Duplicate answers were removed and only one instance was kept to increase coverage. The best 5 answers were output. Answer reranking is the main topic of this paper. Section 4 presents these methods in detail.

4 Dictionary and Web-Based Answer Reranking

4.1 Definition Questions

Compared to other question types, definition questions are special. They are typically very short and in the form of "*What is|are (a|an) X*?", where *X* is a 1 to 3 words term[1], for example: "*What is autism?*", "*What is spider veins?*" and "*What is bangers and mash?*". As we learned from past TREC experience, it was more difficult to find relevant documents for short queries. As stated earlier, over 20% of questions in TREC-10 were of definition type, which was a reflection of real user queries mined from the web search engine logs (Voorhees 2001). Several top performing systems in the evaluation treated this type of question as a special category and most of them used definition answer patterns. The best performing system, InsightSoft-M, (Soubbotin and Soubbotin 2001) used a set of six definition patterns including P1:{<*Q*; is/are; [a/an/the]; *A*>, <*A*; is/are; [a/an/the]; *Q*>} and P2:{<*Q*; comma; [a/an/the]; *A*; [comma/period]>, <*A*; comma; [a/an/the]; *Q*; [comma/period]>}, where *Q* is the term to be defined and *A* is the candidate answer. The InsightSoft-M system returned 88 correct responses based on these patterns. The runner up system (Harabagiu et al. 2001) used 12 answer patterns with extension of WordNet hypernyms. They did not report their success rate for TREC-10 but according to Paşca (2001)[2], this set of patterns with WordNet extension extracted 59 out of 67 definition questions in TREC-8 and TREC-9.

The success stories of these systems indicated that carefully crafted answer patterns were effective in candidate answer extraction. However, just applying answer patterns blindly might lead to disastrous results, as shown by Hermjakob (2002), since correct and incorrect answers were equally likely to match these patterns. For example, for the question "*What is autism?*", the following answers are found in the TREC-10 corpus using the patterns described by the InsightSoft-M system:

① autism$_Q$, a nourishing$_A$, equivocal …
② autism$_Q$, the disorder is$_A$, in fact, …
③ autism$_Q$, the discovery could open new approaches for treating t$_A$he …
④ autism$_Q$ is a mental disorder that is a "severely incapacitatin$_A$g …
⑤ autism$_Q$, the inability to communicate with others$_A$.

Obviously, patterns alone cannot distinguish which one is the best answer. Some other mechanisms are necessary. We propose two different methods to solve this problem. One is a dictionary-based method using WordNet glosses and the other is to go directly to the web and compile web glosses on the fly to help select the best answers. The effect of combining both methods was also studied. We describe these two methods in the following sections.

4.2 Dictionary-Based Reranking

Using a dictionary to look up the definition of a term is the most straightforward solution for answering definition questions. For example, the definition of *autism* in the WordNet is: "*an abnormal absorption with the self; marked by communication disorders and short attention span and inability to treat others as people*". However, we need to find a candidate answer string from the TREC-10 corpus that is equivalent to this definition. By inspection, we find that candidate answers ②, ④, and ⑤ shown in the previous section are more compatible to the definition and ⑤ seems to be the best one.

To automate the decision process, we construct a definition database based on the WordNet noun

[1] Among the 102 TREC-10 definition questions, 81 asked the definition of one word; 19, two words; 2, three words.

[2] Among them 31 were extracted through pattern matching and 27 were from WordNet hypernym expansion.

.

Autism	Webclopedia	Webclopedia + WordNet
1	- Down's syndrome	+ the inability to communicate with others
2	- mental retardation	+ a mental disorder
3	+ the inability to communicate with others	- NIL
4	- NIL	- Down's syndrome
5	- a group of similar-looking diseases	- mental retardation

Table 1. Top 5 answers returned before (Webclopedia) and after (Webclopedia + WordNet) dictionary-based answer reranking for question "*What is autism?*". A "-" indicates wrong answers and a "+" indicates correct answers.

glosses. Closed class words are thrown away and each word w_i in the glosses is assigned a gloss weight s_i^{wn} as follows[3]:

$$s_i^{wn} = \log(N / n_i + 1)$$

where n_i is the number of times word w_i occurring in the WordNet noun glosses and N is total number of occurrences of all noun gloss words in the WordNet. The goodness of the matching M_{wn} for each candidate answer is simply the sum of the weight of the matched word stems between its WordNet definition and itself. For example, candidate answer ⑤ and autism's WordNet definition have these matches: {$inability_5 \Leftrightarrow inability_{wn}$, $communicate_5 \Leftrightarrow communication_{wn}$, $others_5 \Leftrightarrow others_{wn}$}. The reranking score S_{wn} for each candidate answer is its original score multiplied by M_{wn}. The final ranking is then sorted according to S_{wn}, duplicate answers are removed, and the top 5 answers are output. Table 1 shows the top 5 answers returned before and after applying dictionary-based reranking. It demonstrates that dictionary-based reranking not only pushes the best answer to the first place but also boosts other lower ranked good answers i.e. "*a mental disorder*" to the second place.

Harabagiu et al. (2001) also used WordNet to assist in answering definition questions. However, they took the hypernyms of the term to be defined as the default answers while we used its glosses. The hypernym of "*autism*" is "*syndrome*". In this case it would not boost the desired answer to the top but it would instead "validate" "*Down's syndrome*" as a good answer. Further research is needed to investigate the tradeoff between using hypernyms and glosses.

WordNet glosses were incorporated in IBM's statistical question answering system as definition features (Ittycheriah et al. 2001). However, they did not report the effectiveness of the features in definition answer extraction.

Out of vocabulary words is the major problem of dictionary-based reranking. For example, no WordNet entry is found for "*e-coli*" but searching the term "*e-coli*" at www.altavista.com and www.google.com yield the following:

- **E. coli** is a food borne illness. Learn about prevention, symptoms and risks, detection, ... Risks Detection Recent Outbreaks Resources The term ***E. coli*** *is an abbreviation for the bacteria Escherichia*. (1st hit, www.altavista.com)
- The **E. coli** Index (part of the WWW Virtual Library) – Description: Guide to information relating to the *model organism Escherichia coli*. From the WWW Virtual Library. (1st hit, www.google.com)

This brings us to the web-based reranking method that we introduce in the next section.

4.3 Web-Based Reranking

The World Wide Web contains massive amounts of information covering almost any thinkable topic. The TREC-10 questions are typical instances of queries for which users tend to believe answers can be found from the web. However, the candidate answers extracted from the web have to find support in the TREC-10 corpus in order to be judged as correct otherwise they will be marked as unsupported.

The search results of "*e-coli*" from two online search engines indicate that "*e-coli*" is an abbreviation for the bacteria *Escherichia*. However, to automatically identify "*e-coli*" as "*Escherichia*" from these two pages is the same QA problem that we set off to resolve. The only advantage of using the web instead of just the TREC-10 corpus is the assumption that the web contains many more redundant candidate answers due to its huge size. Compared to

[3] This is essentially inverse document (WordNet gloss entry) frequency (IDF) used in the information retrieval research.

.

Wimbledon	Webclopedia	Webclopedia + Google (T=5, W=10, R=70)
1	- the French Open and the U.S. Open.	+ the most famous front yard in tennis and scene
2	- SW20, which includes a Japanese-style water garden	- the French Open and the U.S. Open.
3	+ the most famous front yard in tennis and scene	- NIL
4	- NIL	- Sampras' biggest letdown of the year
5	- Sampras' biggest letdown of the year	- Lawn Tennis & Croquet Club, home of the Wimbledon

Table 2. Top 5 answers returned before (Webclopedia) and after (Webclopedia + Google) web-based answer reranking for question "*What is Wimbledon?*". A "-" indicates wrong answers and a "+" indicates correct answers.

Google's 2,073,418,204 web pages[4], TREC-10 corpus contains *only* about 979,000 articles.

For a given question, we first query the web, apply answer extraction algorithms over a set of top ranked web pages (usually in the lower hundreds), and then rank candidate answers according to their frequency in the set. This assumes the more a candidate answer occurs in the set the more likely it is the correct answer. Clarke et al. (2001) and Brill et al. (2001) both applied this principle and achieved good results.

Instead of using Webclopedia to extract candidate answers from the web and then project back to the TREC-10 corpus, we treat the web as a huge dynamic dictionary. We compile web glosses on the fly for each definition question and apply the same reranking procedure used in the dictionary-based method. We detail the procedure in the following.

(1) Query a search engine (e.g., Altavista) with the term (e.g., "*e-coli*") to be defined.

(2) Download the first R pages (e.g., R = 70).

(3) Extract context word w_i^c within a window of W (e.g., W = 10) words centered at the term to be defined from each page. Closed class words are ignored. These context words are used as candidate web glosses.

(4) The gloss weight s_i^{web} for each word w_i^c is computed as follows[5]:

$$s_i^{web} = t_i \bullet \log(N / n_i + 1)$$

where t_i is the frequency of w_i^c in the set of context words extracted in (3), N is the total number of training questions, and n_i is the number of training questions in which w_i^c occurs. (5) The goodness of the matching M_{web} for each candidate answer is simply the sum of the weights of the matched word stems between its web gloss definition and itself. Only words with gloss weight $s_i^{web} \geq T$ are used to compute M_{web}. The value of T serves as a cut-off threshold to filter out low confidence words.

(6) The reranking score S_{web} for each candidate answer is its original score multiplied by M_{web}. The final ranking is then sorted according to S_{web}, duplicate answers are removed, and the top 5 answers are output. Table 2 shows the top 5 answers returned before and after applying web-based reranking for the question "*What is Wimbledon?*". Google was used as the search engine with T=5, W=10, and R=70.

5 Experiments and Results

We used a set of 102 definition questions from TREC-10 QA track as our test set. The performance of Webclopedia without dictionary or web-based answer reranking was used as the baseline. Webclopedia with dictionary-based answer reranking.

To study the effect of using different search engines, context window sizes, number of top ranked web pages, and web gloss weight cut-off threshold on the performance of web-based answer reranking, we had the following setup:

- Three search engines (E): Altavista (E_A), Google (E_G), and MSNSearch (E_M).
- A run that combined all three search engines' results (E_X).
- Two different context window sizes (W): 5 (W_5) and 10 (W_{10}).
- Eleven sets of top ranked web pages (R_x): top 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100.
- Two different gloss weight cut-off thresholds (T): 5 (T_5) and 10 (T_{10}).

To investigate the performance of combining dictionary and web-based answer reranking, we ran the above setup again but each question's reranking score S_{web+wn} was the multiplication of

[4] This was the number that Google (www.google.com) advertised at its front page as of January 31, 2002.

[5] This is essentially TFIDF (product of term frequency and inverse document frequency) used in the information retrieval research.

.

	W10R10	W10R30	W10R50	W10R70	W10R90
AT5	(0.485,0.637)	(0.499,0.637)	(0.516,0.637)	(0.525,0.667)	(0.519,0.647)
AT10	(0.463,0.618)	(0.497,0.637)	(0.506,0.627)	(0.511,0.657)	(0.502,0.647)
A+T5	(0.503,0.667)	(0.518,0.667)	(0.538,0.676)	(0.555,0.696)	(0.548,0.696)
A+T10	(0.502,0.667)	(0.515,0.667)	(0.528,0.667)	(0.528,0.676)	(0.525,0.676)
GT5	(0.513,0.637)	(0.515,0.647)	(0.536,0.647)	(0.539,0.676)	(0.515,0.657)
GT10	(0.497,0.637)	(0.503,0.647)	(0.527,0.647)	(0.523,0.657)	(0.518,0.637)
G+T5	(0.551,0.686)	(0.537,0.667)	(0.557,0.676)	(0.561,0.725)	(0.547,0.706)
G+T10	(0.536,0.676)	(0.530,0.676)	(0.547,0.676)	(0.544,0.706)	(0.545,0.686)
MT5	(0.521,0.647)	(0.513,0.627)	(0.517,0.647)	(0.514,0.637)	(0.499,0.637)
MT10	(0.505,0.627)	(0.499,0.608)	(0.502,0.637)	(0.488,0.627)	(0.493,0.608)
M+T5	(0.543,0.676)	(0.552,0.667)	(0.544,0.676)	(0.542,0.696)	(0.533,0.676)
M+T10	(0.527,0.647)	(0.537,0.647)	(0.525,0.667)	(0.519,0.696)	(0.520,0.667)
XT5	(0.526,0.647)	(0.539,0.676)	(0.533,0.627)	(0.519,0.637)*	(0.515,0.627)
XT10	(0.509,0.618)	(0.524,0.657)	(0.532,0.627)	(0.524,0.647)	(0.517,0.637)
X+T5	(0.553,0.696)	(0.551,0.696)	(0.556,0.686)	(0.550,0.696)	(0.546,0.686)
X+T10	(0.531,0.657)	(0.543,0.686)	(0.555,0.686)	(0.550,0.696)	(0.546,0.686)

Table 3. Results of 90 runs shown in (MRR, PCT5) score pair where A: Altavista, G: Google, M: MSNSearch, X: all three search engines, W: context window size, R: number of top ranked web paged used, T: web gloss weight cut-off threshold. Runs marked with '+' indicate both dictionary and web-based answer reranking are used.

its original score, web-based matching score M_{web}, and dictionary-based matching score M_{wn}. A total of 354 runs were performed. Manual evaluation of these 354 runs was not impossible but would be time consuming. We instead used the answer patterns provided by NIST to score all runs automatically.

Due to space constraint, Table 3 shows the (MRR, PCT5) score pair for 90 runs out of 352 runs. The other two runs were the baseline run with a score pair of (0.450, 0.637) and the dictionary-based run, (0.535, 0.667). The best run was the combined dictionary and web-based run using Google as the search engine with 10-word context window, 70 top ranked pages, and a gloss weight cut-off threshold of 5. Analyzing all runs according to Table 3, we made the following observations.

(1) Dictionary-based reranking improved baseline performance by 19% in MRR and 5% in PCT5 (MRR: 0.535, PCT5: 0.667).

(2) The best web-based reranking (MRR: 0.539, PCT5: 0.676) was achieved with W=10, R=70, and T=5. It was comparable to the dictionary-based reranking.

(3) Web-based reranking generally improved results. Only 6 runs[6] (not shown in the table) did worse in their MRR scores than just using Webclopedia alone and these runs concentrated on low ranked page counts of 5 and 10.

(4) Different search engines reached their best performance at different parameter settings. Overall Google did better.

(5) Combining multiple search engine results (runs designed with X and X+) did not always improve performance. In some cases, it even degraded system performance ($E_xT_5W_{10}R_{70}$: 0.519, 0.637).

(6) Lower web gloss weight cut-off threshold was better at 5.

(7) Longer context window was better at 10 (not shown in the table).

(8) Taking top ranked pages of 50 to 90 pages provided better results.

(9) Combining dictionary and web-based reranking always did better than using the web-based method alone.

(10) Using WordNet and Google together was always better than just using WordNet alone in both MRR and PCT5 (the underlined cells).

5.1 Question Difficulty

To investigate the effectiveness of using dictionary and web-based answer reranking on question of different difficulty, we define question difficulty as: $d = 1 - (n/N)$, where n is the number of systems participating in TREC-10 that returned answers in top 5 and N is the number of total runs (that is, 67 for TREC-10). When d = 1 no systems provided an answer in top 5; while d = 0 if all runs provided at least one answer in top 5. Table 4 shows the improvement of MRR and PCT5 scores at four different question difficulty levels with four different system setups. The results indicate that using either dictionary or web-based answer reranking improved system performance at all levels. The best results were achieved when evidence from both resources was used. However, it also demonstrates the difficulty of improving performance on very hard questions (d>=0.75). This implies we might need to consider alternative methods to improve the system performance further.

	d>=0.00 (102)	d>=0.25 (95)	d>=0.50 (71)	d>=0.75 (40)
F	(0.450,0.637)	(0.394,0.611)	(0.264,0.549)	(0.084,0.375)
F+	(0.535,0.667)	(0.474,0.642)	(0.323,0.592)	(0.100,0.375)
FG	(0.539,0.676)	(0.475,0.653)	(0.319,0.592)	(0.125,0.375)
F+G	(0.561,0.725)	(0.498,0.705)	(0.333,0.648)	(0.128,0.400)

Table 4. System performance at different question difficulty levels. (F: Webclopedia only, F+: Webclopedia with WordNet, FG: Webclopedia with Google, and F+G: Webclopedia with

[6] These were $E_AT_5W_5R_5$ (0.437, 0.598), $E_AT_{10}W_5R_5$ (0.434, 0.608), $E_AT_{10}W_{10}R_5$ (0.437, 0.598), $E_MT_5W_5R_5$ (0.436, 0.608), $E_MT_{10}W_5R_5$ (0.438, 0.608), and $E_MT_{10}W_{10}R_5$ (0.443, 0.618).

.

6 Conclusions

We described dictionary-based answer reranking using WordNet, web-based answer reranking using three different online search engines, and their evaluations at various parameter settings on a set of 102 TREC-10 definition questions. We showed that using either approach alone improved MRR score by 19% and PCT5 score by 5% over the baseline. However, the best performance was achieved when both methods were used together. In that setting a 25% increase in MRR score and 14% improvement in PCT5 score were obtained.

The difference on the best MRR and PCT5 scores (0.56 vs. 0.73) suggests neither dictionary-based nor web-based will solve the reranking problem completely.

To improve the performance further, we need better ways to compile web glosses and combine them with WordNet glosses. We also need a better combination function — a statistical model for combining patterns, dictionary, and web scores. We have started investigating the possibility of applying answer reranking to other question types and exploring specialized web resources.

References

Abney, S., M. Collins, and A. Singhal. 2000. Answer Extraction. In *Proceedings of the Applied Natural Language Processing Conference* (ANLP-NAACL-00), Seattle, WA, 296–301.

Bikel, D., R. Schwartz, and R. Weischedel. 1999. An Algorithm that Learns What's in a Name. In *Machine Learning—Special Issue on NL Learning*, 34, 1–3.

Brill, E., J. Lin, M. Banko, S. Dumais, and A. Ng. 2001. Data-Intensive Question Answering. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 183–189.

Buchholz, S. 2001. Using Grammatical Relations, Answer Frequencies and the World Wide Web for TREC Question Answering. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 496–503.

Clarke, C.L.A., G.V. Cormack, T.R. Lynam, C.M. Li, and G.L. McLearn. 2001. Web Reinforced Question Answering. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 620–626.

Fellbaum, Ch. (ed). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Fukumoto, J, T. Kato, and F. Masui. 2001. NTCIR Workshop 3 QA Task – Question Answering Challenge (QAC). http://research.nii.ac.jp/ntcir/ workshop/qac/cfp-en.html.

Harabagiu, S., D. Moldovan, M. Paşca, R. Mihalcea, M. Surdeanu, R. Buneascu, R. Gîrju, V. Rus and P. Morarescu. 2001. FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the 9th Text Retrieval Conference* (TREC-9), NIST, 479–488.

Hermjakob, U. 2001. Parsing and Question Classification for Question Answering. In *Proceedings of the Workshop on Open-Domain Question Answering* post-conference workshop of ACL-2001, Toulouse, France.

Hermjakob, U. 2002. Open Questions Regarding Precision of the Insight Q&A System. *Personal communication.*.

Hovy, E.H., U. Hermjakob, and C.-Y. Lin. 2001. The Use of External Knowledge in Factoid QA. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 166–174.

E.H. Hovy, U. Hermjakob, C-Y. Lin, and D. Ravichandran. 2002. Using Knowledge to Facilitate Pinpointing of Factoid Answers. In *Proceedings of the 19th International Conference on Computational Linguistics* (COLING 2002), Taipei, Taiwan, August 24 - September 1, 2002.

Ittycheriah, A., M. Franz, and S. Roukos. 2001. IBM's Statistical Question Answering System. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 317–323.

Paşca, M. 2001. *High Performance, Open-Domain Question Answering from Large Text Collections*. Ph.D. dissertation, Southern Methodist University, Dallas, TX.

Soubbotin, M.M. and S.M. Soubbotin. 2001. Patterns of Potential Answer Expressions as Clues to the Right Answer. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 175–182.

Srihari, R. and W. Li. 2000. A Question Answering System Supported by Information Extraction. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (ANLP-NAACL-00), Seattle, WA, 166–172.

Voorhees, E.M. 1999. The TREC-8 Question and Answering Track Report. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 77–82, 2000. NIST Special Publication 500-246.

Voorhees, E. 2001. Overview of the Question Answering Track. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 71–81.

Voorhees, E.M. and D.K. Harman. 2001. The Proceedings of the Eighth Text REtrieval Conference (TREC-10), NIST.

Witten, I.H., A. Moffat, and T.C. Bell. 1994. *Managing Gigabytes: Compressing and Indexing Documents and Images*. New York: Van Nostrand Reinhold.