Using Knowledge to Facilitate Factoid Answer Pinpointing

Eduard Hovy, Ulf Hermjakob, Chin-Yew Lin, Deepak Ravichandran
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
USA
{hovy,ulf,cyl,ravichan}@isi.edu

Abstract

In order to answer factoid questions, the Webclopedia QA system employs a range of knowledge resources. These include a QA Typology with answer patterns, WordNet, information about typical numerical answer ranges, and semantic relations identified by a robust parser, to filter out likely-looking but wrong candidate answers. This paper describes the knowledge resources and their impact on system performance.

1. Introduction

The TREC evaluations of QA systems (Voorhees, 1999) require answers to be drawn from a given source corpus. Early QA systems used a simple filtering technique, question word density within a fixed *n*-word window, to pinpoint answers. Robust though this may be, the window method is not accurate enough. In response, factoid question answering systems have evolved into two types:

- Use-Knowledge: extract query words from the input question, perform IR against the source corpus, possibly segment resulting documents, identify a set of segments containing likely answers, apply a set of heuristics that each consults a different source of knowledge to score each candidate, rank them, and select the best (Harabagiu et al., 2001; Hovy et al., 2001; Srihari and Li, 2000; Abney et al., 2000).
- Use-the-Web: extract query words from the question, perform IR against the web, extract likely answer-bearing sentences, canonicalize the results, and select the most frequent answer(s). Then, for justification, locate examples of the answers in the source corpus (Brill et al., 2001; Buchholz, 2001).

Of course, these techniques can be combined: the popularity ratings from Use-the-Web can also be applied as a filtering criterion (Clarke et al., 2001), or the knowledge resource heuristics can filter the web results. However, simply going to the web without using further knowledge (Brill et al., 2001) may return the web's majority opinions on astrology, the killers of JFK, the cancerous effects of microwave ovens, etc.—fun but not altogether trustworthy.

In this paper we describe the range of filtering techniques our system Webclopedia applies, from simplest to most sophisticated, and indicate their impact on the system.

2. Webclopedia Architecture

As shown in Figure 1, Webclopedia adopts the Use-Knowledge architecture. Its modules are described in more detail in (Hovy et al., 2001; Hovy et al., 1999):

- **Question parsing:** Using BBN's IdentiFinder (Bikel et al., 1999), the CONTEX parser (Hermjakob, 1997) produces a syntactic-semantic analysis of the question and determines the QA type.
- **Query formation**: Single- and multi-word units (content words) are extracted from the analysis, and WordNet synsets (Fellbaum, 1998) are used for query expansion. A series of Boolean queries of decreasing specificity is formed.
- **IR**: The publicly available IR engine MG (Witten et al., 1994) returns the top-ranked *N* documents.
- **Selecting and ranking sentences**: For each document, the most promising *K* sentences are located and scored using a formula that

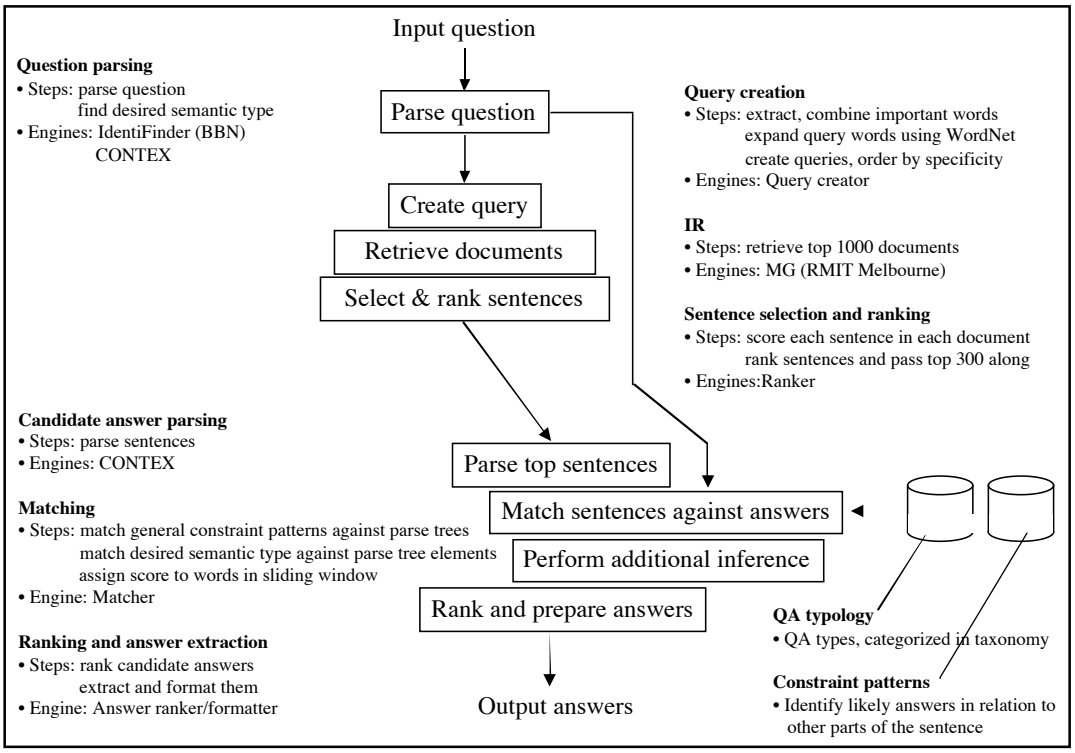


Figure 1. Webclopedia architecture.

rewards word and phrase overlap with the question and its expanded query words. Results are ranked.

- **Parsing candidates**: CONTEX parses the top-ranked 300 sentences.
- **Pinpointing**: As described in Section 3, a number of knowledge resources are used to perform filtering/pinpointing operations.
- **Ranking of answers**: The candidate answers' scores are compared and the winner(s) are output.

3. Knowledge Used for Pinpointing

3.1 Type 1: Question Word Matching

Unlike (Prager et al., 1999), we do not first annotate the source corpus, but perform IR directly on the source text, using MG (Witten et al., 1994). To determine goodness, we assign an initial base score to each retrieved sentence. We then compare the sentence to the question and adapt this score as follows:

- exact matches of proper names double the base score.
- matching an upper-cased term adds a 60% bonus of the base score for multi-words terms and 30% for single words (matching "United States" is better than just "United").
- matching a WordNet synonym of a term discounts by 10% (lower case) and 50% (upper case). (When "Cage" matches "cage", the former may be the last name of a person and the latter an object; the case mismatch signals less reliability.)
- lower-case term matches after Porter stemming are discounted 30%; upper-case matches 70% (Porter stemming is more aggressive than WordNet stemming).
- Porter stemmer matches of both question and sentence words with lower case are discounted 60%; with upper case, 80%.
- if CONTEX indicates a term as being *qsubsumed* (see Section 3.9) the term is discouned 90% (in "Which country manufactures weapons of mass destruction?", "country" will be marked as *qsubsumed).*

The top-scoring 300 sentences are passed on for further filtering.

3.2 Type 2: Qtargets, the QA Typology, and the Semantic Ontology

We classify desired answers by their semantic type, which have been taxonomized in the Webclopedia QA Typology (Hovy et al., 2002), http://www.isi.edu/natural-language/projects/webclopedia/Taxonomy/taxonomy_toplevel.html). The currently approx. 140 classes, which we call *qtargets*, were developed after an analysis of over 17,000 questions (downloaded in 1999 from answers.com) and later enhancements to Webclopedia. They are of several types:

- common semantic classes such as PROPER-PERSON, EMAIL-ADDRESS, LOCATION, PROPER-ORGANIZATION;
- classes particular to QA such as YES:NO, ABBREVIATION-EXPANSION, and WHY-FAMOUS;
- syntactic classes such as NP and NOUN, when no semnatic type can be determined (e.g., "What does Peugeot manufacture?");
- roles and slots, such as INSTRUMENT-OF and COLOR-OF respectively, to indicate a desired relation with an anchoring concept.

Given a question, the CONTEX parser uses a set of 276 hand-built rules to identify its most likely qtarget(s), and records them in a backoff scheme (allowing more general qtarget nodes to apply when more specific ones fail to find a match). The generalizations are captured in a typical concept ontology, a 10,000-node extract of WordNet.

The recursive part of pattern matching is driven mostly by interrogative phrases. For example, the rule that determines the applicability of the qtarget WHY-FAMOUS requires the question word "who", followed by the copula, followed by a proper name. When there is no match at the current level, the system examines any interrogative constituent, or words in special relations to it. For example, the qtarget TEMPERATURE-QUANTITY (as in "What is the melting point of X?" requires as syntactic object something that in the ontology is subordinate to TEMP-QUANTIFIABLE-ABS-TRACT with, as well, the word "how" paired with "warm", "cold", "hot", etc., or the phrase "how many degrees" and a TEMPERATURE-UNIT (as defined in the ontology).

3.3 Type 3: Surface Pattern Matching

Often qtarget answers are expressed using rather stereotypical words or phrases. For example, the year of birth of a person is typically expressed using one of these phrases:

<name> was born in <birthyear>
<name> (<birthyear>–<deathyear>)

We have developed a method to learn such patterns automatically from text on the web (Ravichandran and Hovy, 2002). We have added into the QA Typology the patterns for appropriate qtargets (qtargets with closed-list answers, such as PLANETS, require no patterns). Where some QA systems use such patterns exclusively (Soubbotin and Soubbotin, 2001) or partially (Wang et al., 2001; Lee et al., 2001), we employ them as an additional source of evidence for the answer. Preliminary results on for a range of qtargets, using the TREC-10 questions and the TREC corpus, are:

Question type (qtarget)	Number of questions	MRR on TREC docs
BIRTHYEAR	8	0.47875
INVENTORS	6	0.16667
DISCOVERERS	4	0.1250
DEFINITIONS	102	0.3445
WHY-FAMOUS	3	0.6666
LOCATIONS	16	0.75

3.4 Type 4: Expected Numerical Ranges

Quantity-targeting questions are often underspecified and rely on culturally shared cooperativeness rules and/or world knowledge:

Q: How many people live in Chile?

S1: "From our correspondent comes good news about the nine people living in Chile…"

A1: nine

While certainly nine people do live in Chile, we know what the questioner intends. We have hand-implemented a rule that provides default range assumptions for POPULATION questions and biases quantity questions accordingly.

3.5 Type 5: Abbreviation Expansion

Abbreviations often follow a pattern:

Q: What does NAFTA stand for?

S1: "This range of topics includes the North American Free Trade Agreement, NAFTA, and the world trade agreement GATT."

S2: "The interview now changed to the subject of trade and pending economic issues, such as the issue of opening the rice market, NAFTA, and the issue of Russia repaying economic cooperation funds."

After Webclopedia identifies the qtarget as ABBREVIATION-EXPANSION, it extracts possible answer candidates, including "North American Free Trade Agreement" from S1 and "the rice market" from S2. Rules for acronym matching easily prefer the former.

3.6 Type 6: Semantic Type Matching

Phone numbers, zip codes, email addresses, URLs, and different types of quantities obey lexicographic patterns that can be exploited for matching, as in

Q: What is the zip code for Fremont, CA?

S1: "…from Everex Systems Inc., 48431 Milmont Drive, Fremont, CA 94538."

and

Q: How hot is the core of the earth?

S1. "The temperature of Earth's inner core may be as high as 9,000 degrees Fahrenheit (5,000 degrees Celsius)."

Webclopedia identifies the qtargets respectively as ZIP-CODE and TEMPERATURE-QUANTITY. Approx. 30 heuristics (cascaded) apply to the input before parsing to mark up numbers and other orthographically recognizable units of all kinds, including (likely) zip codes, quotations, year ranges, phone numbers, dates, times, scores, cardinal and ordinal numbers, etc. Similar work is reported in (Kwok et al., 2001).

3.7 Type 7: Definitions from WordNet

We have found a 10% increase in accuracy in answering definition questions by using external glosses obtained from WordNet. For

Q: What is the Milky Way?

Webclopedia identified two leading answer candidates:

A1: outer regions

A2: the galaxy that contains the Earth

Comparing these with the WordNet gloss:

WordNet: "Milky Way—the galaxy containing the solar system"

allows Webclopedia to straightforwardly match the candidate with the greater word overlap.

Curiously, the system also needs to use WordNet to answer questions involving common knowledge, as in:

Q: What is the capital of the United States?

because authors of the TREC collection do not find it necessary to explain what Washington is:

Ex: "Later in the day, the president returned to Washington, the capital of the United States."

While WordNet's definition

Wordnet: "Washington—the capital of the United States"

directly provides the answer to the matcher, it also allows the IR module to focus its search on passages containing "Washington", "capital", and "United States", and the matcher to pick a good motivating passage in the source corpus.

Clearly, this capability can be extended to include (definitional and other) information provided by other sources, including encyclopedias and the web.

3.8 Type 8: Semantic Relation Matching

So far, we have considered individual words and groups of words. But often this is insufficient to accurately score an answer. As also noted in (Buchholz, 2001), pinpointing can be improved significantly by matching semantic relations among constituents:

Q: Who killed Lee Harvey Oswald?

Qtargets: PROPER-PERSON & PROPER-NAME, PROPER-ORGANIZATION

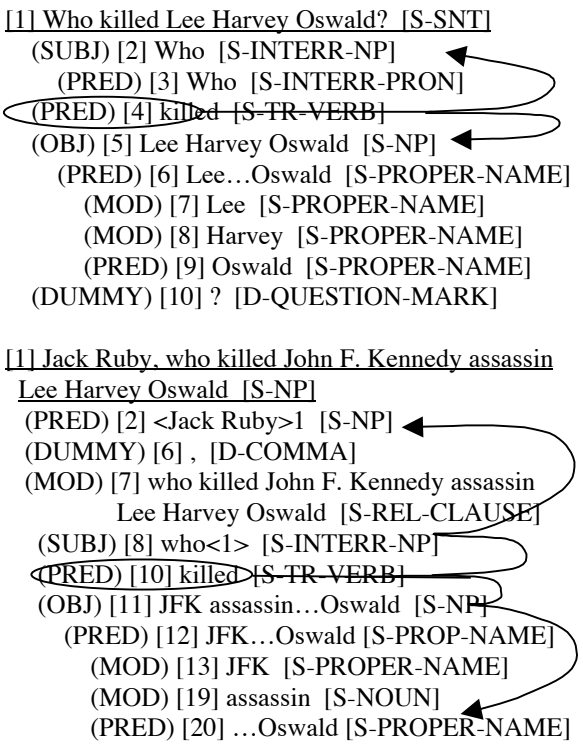
S1: "Belli's clients have included **Jack Ruby**, who killed John F. Kennedy assassin Lee Harvey Oswald, and Jim and Tammy Bakker."

S2: "On Nov. 22, 1963, the building gained national notoriety when Lee Harvey Oswald allegedly shot and killed **President John F. Kennedy** from a sixth floor window as the presidential motorcade passed."

The CONTEX parser (Hermjakob, 1997; 2001) provides the semantic relations. The parser uses machine learning techniques to build a robust grammar that produces semantically annotated syntax parses of English (and Korean and Chinese) sentences at approx. 90% accuracy (Hermjakob, 1999).

The matcher compares the parse trees of S1 and S2 to that of the question. Both S1 and S2 receive credit for matching question words "Lee Harvey Oswald" and "kill" (underlined), as well

as for finding an answer (bold) of the proper qtarget type (PROPER-PERSON). However, is the answer "Jack Ruby" or "President John F. Kennedy"? The only way to determine this is to consider the semantic relationship between these candidates and the verb "kill" (parse trees simplified, and only portions shown here):



Although the PREDs of both S1 and S2 match that of the question "killed", only S1 matches "Lee Harvey Oswald" as the head of the logical OBJect. Thus for S1, the matcher awards additional credit to node [2] (Jack Ruby) for being the logical SUBJect of the killing (using anaphora resolution). In S2, the parse tree correctly records that node [13] ("John F. Kennedy") is not the object of the killing. Thus despite its being closer to "killed", the candidate in S2 receives no extra credit from semantic relation matching.

It is important to note that the matcher awards extra credit for *each* matching semantic relationship between two constituents, not only when everything matches. This granularity improves robustness in the case of partial matches.

Semantic relation matching applies not only to logical subjects and objects, but also to all other roles such as location, time, reason, etc. (for additional examples see http://www.isi.edu/natural-language/projects/webclopedia/sem-rel-examples.html). It also applies at not only the sentential level, but at all levels, such as post-modifying prepositional and pre-modifying determiner phrases

Additionally, Webclopedia uses 10 lists of word variations with a total of 4029 entries for semantically related concepts such as "to invent", "invention" and "inventor", and rules for handling them. For example, via coercing "invention" to "invent", the system can give "Johan Vaaler" extra credit for being a likely logical subject of "invention":

Q: Who invented the paper clip?

Qtargets: PROPER-PERSON & PROPER-NAME, PROPER-ORGANIZATION

S1: "The paper clip, weighing a desk-crushing 1,320 pounds, is a faithful copy of **Norwegian Johan Vaaler**'s 1899 invention, said Per Langaker of the Norwegian School of Management."

while "David" actually loses points for being outside of the clausal scope of the inventing:

S2: "'Like the guy who invented the safety pin, or the guy who invented the paper clip,' **David** added."

3.9 Type 9: Word Window Scoring

Webclopedia also includes a typical window-based scoring module that moves a window over the text and assigns a score to each window position depending on a variety of criteria (Hovy et al., 1999). Unlike (Clarke et al., 2001; Lee et al., 2001; Chen et al., 2001), we have not developed a very sophisticated scoring function, preferring to focus on the modules that employ information deeper than the word level.

This method is applied only when no other method provides a sufficiently high-scoring answer. The window scoring function is

$S = (500/(500+w))*(1/r) * \square[(\square I^{1.5*q*e*b*u})^{1.5}]$

Factors:

w: window width (modulated by gaps of various lengths: "white house" ≠ "white car and house"),

r: rank of qtarget in list returned by CONTEX,

I: window word information content (inverse log frequency score of each word), summed,

q: # different question words matched, plus specific rewards (bonus *q=3.0*),

e: penalty if word matches one of question word's WordNet synset items (*e=0.8*),

b: bonus for matching main verb, proper names, certain target words (*b=2.0*),

u: (value 0 or 1) indicates whether a word has been qsubsumed ("subsumed" by the qtarget) and should not contribute (again) to the score. For example, "In what year did Columbus discover America?" the qsubsumed words are "what" and "year".

4. Performance Evaluation

In TREC-10's QA track, Webclopedia received an overall Mean Reciprocal Rank (MRR) score of 0.435, which put it among the top 4 performers of the 68 entrants (the average MRR score for the main QA task was about 0.234). The pinpointing heuristics are fairly accurate: when Webclopedia finds answers, it usually ranks them in the first place (1st place: 35.5%; 2nd: 8.94%; 3rd: 5.69%; 4th: 3.05%; 5th: 5.28%; not found: 41.87%).

We determined the impact of each knowledge source on system performance, using the TREC-10 test corpus using the standard MRR scoring. We applied the system to the questions of each knowledge type separately, with and without its specific knowledge source/algorithm. Results are shown in Table 1, columns A (without) and B (with). To indicate overall effect, we also show (in columns C and D) the percentage of questions in TREC-10 and -9 respecively of each knowledge type.

5. Conclusions

It is tempting to search for a single technique that will solve the whole problem (for example, Ittycheriah et al. (2001) focus on the subset of factoid questions answerable by NPs, and train a statistical model to perform NP-oriented answer pinpointing). Our experience, however, is that even factoid QA is varied enough to require various special-purpose techniques and knowledge. The theoretical limits of the various techniques are not known, though Light et al.'s (2001) interesting work begins to study this.

Column A: % questions of the knowledge type answered correctly without using knowlege
Column B: % questions, now using knowledge
Column C: % questions of type in TREC-10
Column D: % questions of type in TREC-9

	A	B	C	D
Abbreviation exp.	20.0	70.0	1.0	2.3
Number ranges	50.0	50.0	1.2	1.8
WordNet (def Qs)	48.3	67.5	20.9	5.1
Semantic types				
- locator types	N/A	N/A	0.0	0.4
- quantity types	22.5	48.7	10.8	5.5
- date/year types	45.0	57.3	9.2	10.2
Patterns				
- definitions	–	34.4	20.9	5.1
- why-famous	–	66.7	0.6	–
- locations	–	75.0	3.2	–
- birthyear	–	47.9	1.6	–
Semantic relations	39.4	46.5	72.2	85.7

Table 1. Performance of knowledge sources. Semantic relation scores measured only on questions in which they could logically apply.

We conclude that factoid QA performance can be significantly improved by the use of knowledge attuned to specific question types and specific information characteristics. Most of the techniques for exploiting this knowledge require learning to ensure robustness. To improve performance beyond this, we believe a combination of going to the web and turning to deeper world knowledge and automated inference (Harabagiu et al., 2001) to be the answer. It remains an open question how much work these techniques would require, and what their payoff limits are.

References

Abney, S., M. Collins, and A. Singhal. 2000. Answer Extraction. *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. *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. *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. *Proceedings of the TREC-10 Conference*. NIST, 496–503.

Chen, J., A.R. Diekema, M.D. Taffet, N. McCracken, N. Ercan Ozgencil, O. Yilmazel, and E.D. Liddy. 2001. CNLP at TREC-10 QA Track. *Proceedings of the TREC-10 Conference*. NIST, 480–490.

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

Clarke, C.L.A., G.V. Cormack, and T.R. Lynam. 2001. Exploiting Redundancy in Question Answering. *Proceedings of the SIGIR Conference*. New Orleans, LA, 358–365.

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

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

Hermjakob, U. 1997. *Learning Parse and Translation Decisions from Examples with Rich Context*. Ph.D. dissertation, University of Texas Austin. file://ftp.cs.utexas.edu/pub/mooney/papers/hermjakob-dissertation 97.ps.gz.

Hermjakob, U. 2001. Parsing and Question Classification for Question Answering. *Proceedings of the Workshop on Question Answering at ACL-2001.* Toulouse, France.

Hovy, E.H., L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. 1999. Question Answering in Webclopedia. *Proceedings of the TREC-9 Conference*. NIST. Gaithersburg, MD, 655–673.

Hovy, E.H., U. Hermjakob, and D. Ravichandran. 2002. A Question/Answer Typology with Surface Text Patterns. Poster in *Proceedings of the DARPA Human Language Technology Conference (HLT)*. San Diego, CA, 234–238.

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

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

Kwok, K.L., L. Grunfeld, N. Dinstl, and M. Chan. 2001. TREC2001 Question-Answer, Web and Cross Language experiments using PIRCS. *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 447–451.

Lee, G.G., J. Seo, S. Lee, H. Jung, B-H. Cho, C. Lee, B-K. Kwak, J, Cha, D. Kim, J-H. An, H. Kim, and K. Kim. 2001. SiteQ: Engineering High Performance QA System Using Lexico=Semantic Pattern Matching and Shallow NLP. *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 437–446.

Light, M., G.S. Mann, E. Riloff, and E. Breck. 2001. Analyses for Elucidating Current Question Answering Technology. *Natural Language Engineering*, to appear.

Oh, JH., KS. Lee, DS. Chang, CW. Seo, and KS. Choi. 2001. TREC-10 Experiments at KAIST: Batch Filtering and Question Answering. *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, 354–361.

Prager, J., E. Brown, D.R. Radev, and K. Czuba. 1999. One Search Engine or Two for Question Answering. *Proceedings of the TREC-9 Conference*. NIST, Gaithersburg, MD, 235–240.

Ravichandran, D. and E.H. Hovy. 2002. Learning Surface Text Patterns for a Question Answering System. *Proceedings of the ACL conference.* Philadelphia, PA.

Soubbotin, M.M. and S.M. Soubbotin. 2001. Patterns of Potential Answer Expressions as Clues to the Right Answer. *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. *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. 1999. Overview of the Question Answering Track. *Proceedings of the TREC-9 Conference*. NIST, Gaithersburg, MD, 71–81.

Wang, B., H. Xu, Z. Yang, Y. Liu, X. Cheng, D. Bu, and S. Bai. 2001. TREC-10 Experiments at CAS-ICT: Filtering, Web, and QA. *Proceedings of the TREC-10 Conference*. NIST, 229–241.

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