

Web-based Question Answering (panel)

Dragomir R. Radev, University of Michigan (moderator)

Panelists:

Susan Dumais, Microsoft Research

Eduard Hovy, ISI

Boris Katz, MIT

Brian Ulicny, Ask Jeeves

Abstract

Early TREC-style Question Answering Systems were characterized by the following features: (a) the answer of the question was known to be included in a given local corpus, (b) the size of the small corpus permitted preprocessing, including named entity extraction and parsing of all documents, and (c) the corpus consisted of well-written news documents. More recently, QA Systems have started to use the Web as a corpus, either by extracting answers from the Web rather than a local corpus or by learning lexical patterns from the Web which are then used to improve the system itself. Using the Web for Question Answering presents an interesting combination of opportunities and challenges. This panel will discuss how to leverage the opportunities while addressing the inevitable challenges.

The Web as a repository of answers

Scaling QA systems to the Web presents an extraordinary challenge. The collections used in TREC8-10 contained a couple of hundred thousand documents while Google indexes more than 3 Billion Web pages. While preprocessing (POS tagging, chunking, semantic entity annotation) the TREC corpus is feasible, doing the same with Google's index is out of the question with current technology. Furthermore, the TREC corpus is limited to a short and fixed time period while new documents are constantly added to the Web. Web-based question answering systems therefore also need to deal with issues such as redundancy, recency, and reliability. While the Web presents a number of interesting challenges to Question Answering, it is important to observe that a number of existing systems successfully address many of them and achieve high-quality results.

Many of the systems discussed so far use existing search engines such as Google as their backends. Even though

Google's performance at getting documents containing the answers to natural language questions even without making changes to the questions (e.g., see Radev et al. 01), most systems deploy *query modulation* to convert each question to a query in the specific language of the search engine. In addition to query modulation and an interface to public search engines, all systems also employ some form of passage retrieval and passage ranking. To some extent, all systems are architecturally isomorphic. At the same time, all major systems exhibit some interesting properties that set them apart from each other.

MIT's *START* system (Katz 97) has been on line for a number of years and has traditionally focused more on questions about the MIT InfoLab and geography. Recently, it has expanded to include general domain questions.

Ask Jeeves (www.ask.com) is the most popular search engine that allows natural language questions. User questions are then matched against a database of known questions and answers. Ask Jeeves is not, strictly speaking, a Q&A system, given that its results come as documents, rather than answers.

U. Washington's system, *Mulder* (Kwok et al. 01) was the first major Q&A system to exploit the Web as a source of answers. Mulder sends a question to several search engines and uses a natural language parser to extract likely answers.

The *Waterloo* team (Clarke et al. 01) reduce the problem of question answering to a two-stage process: first, they extract relevant passages using IR techniques, then decide what passage is most likely to contain the answer by its frequency.

Microsoft's system (Dumais et al. 02) also looks at redundancy of n-grams to identify good answers.

Michigan's system, *NSIR* (Radev et al. 02), uses probabilistic phrase reranking based on part of speech

patterns to identify the semantic type of a potential answer phrase.

InsightSoft's approach (Soubbotin 01) makes use of automatically extracted patterns to identify likely passages given a question. For example, a pattern like “capitalized word ‘in’ digit digit digit digit ‘born’” is indicative of a date-of-birth question.

Ravichandran and Hovy from *ISI* (Ravichandran and Hovy 02) go to the Web to extract patterns of the style suggested by Soubbotin. They send queries containing known question-answer pairs to the Web and retrieve phrases that are indicative of a relation between the question and the answer words.

While the overall architecture of Web-based Q&A systems is fairly standard, the amount of Natural Language Processing techniques used varies from system to system. An interesting open question therefore is to what extent NLP components help achieve higher performance.

Discussion questions

To ensure a focused discussion, panelists are asked ahead of time to prepare short statements addressing the following questions:

1. What challenges does the Web pose for Q&A when compared to the TREC corpus?
2. What Web-based resources can be used for Q&A?
3. How does crawling/indexing the Web change for Q&A? How much Q&A processing can be done offline (before a question comes in)?
4. What NLP techniques scale up to the Web (e.g., semantic entity annotation, anaphora resolution, statistical parsing)?

Each panelist will have 5-10 minutes to present their statement. A general discussion including all panelists and the audience will follow.

Future Challenges

One of the goals of the panel is to identify some of the challenges facing developers of Web-based Q&A systems and to propose ways to address them. Some problems expected to make the list include Q&A in languages other than English, temporal question answering, and answering questions based on a model of the user.

References

Charles Clarke, Gordon Cormack, and Thomas Lynam. *Exploiting Redundancy in Question Answering*. SIGIR 2001, New Orleans, LA.

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. *Web Question Answering: Is More Always Better?* SIGIR 2003, Tampere, Finland.

Boris Katz. *From Sentence Processing to Information Access on the World Wide Web*. AAAI Spring Symposium on Natural Language Processing for the Web 1997, Stanford, CA.

Cody Kwok, Oren Etzioni, and Daniel Weld. *Scaling Question Answering Systems to the Web*. WWW 2001, Hong Kong.

Dragomir Radev, Kelsey Libner, and Weiguo Fan. *Getting Answers to Natural Language Queries on the Web*. JASIST 53(5), January 2001.

Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, Amardeep Grewal. *Probabilistic Question Answering on the Web*. WWW 2002, Honolulu, HI.

Deepak Ravichandran and Eduard Hovy. *Learning surface text patterns for a Question Answering systems*. ACL 2002, Philadelphia, PA.

M. M. Soubbotin. *Patterns of Potential Answer Expressions as Clues to the Right Answers*. TREC 2001, Gaithersburg, MD.