Aranea: Mining Answers from the World Wide Web

Jimmy Lin & Boris Katz

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Users seeking answers to questions like “what Spanish explorer discovered the Mississippi River” should not have to peruse entire documents about the Spanish explorers or early American history to obtain relevant information. Instead, a computer system should automatically distill large amounts of text into a compact answer in response to a question posed in natural language.

Motivation: An enormous amount of textual data is freely available on the World Wide Web; properly utilized, this vast repository of text can serve as a rich source of knowledge for answering fact-based questions. The sheer size of the Web is its primary asset; items of knowledge are stated multiple times, in multiple documents, and in multiple ways. This unique characteristic, known as data redundancy [1], allows for the development of novel question answering techniques.

Previous Work: Question answering systems have traditionally been viewed as an extension of information retrieval systems (a notable exception is START; see [2] and other abstracts in this volume for details). The general approach has been to couple passage-level retrieval techniques with shallow linguistic processing capabilities to locate short passages likely to contain an answer. [3] Additionally, most question answering systems attempt to extract answers from fixed, static corpora, i.e., encyclopedia articles or newspaper archives.

Approach: Consider two possible answers to the question “Who killed Abraham Lincoln?”

(1) John Wilkes Booth killed Abraham Lincoln.
(2) John Wilkes Booth is perhaps America’s most infamous assassin. He is best known for firing the bullet that ended Abraham Lincoln’s life.

One would surely agree that the answer could be more easily extracted from passage (1) than from passage (2). In general, the task of answering a question is not very difficult if the document collection contains the answer stated as a simple reformulation of the question. As the size of the target document collection grows, the more likely it is that question answering systems can find statements that answer the question in an obvious way.

Surprisingly, the World Wide Web is so big that simple pattern matching and statistical techniques suffice to extract the answer to many fact-based questions, obviating the need to understand both the structure and meaning of language. The answer to a question can be extracted by searching directly for an anticipated answer form, e.g., in the above example, by searching for the string “killed Abraham Lincoln” and extracting words occurring to the left. Naturally, this simple technique depends crucially on the corpus having an answer formulated in a specific way. Thus, the larger the text collection is, the greater the probability that simple pattern matching techniques will yield the correct answer. Fortunately, the Web is large enough such that these techniques are viable. Thus, data redundancy enables us to utilize simple tricks to overcome many troublesome issues in natural language processing, e.g., alternations, anaphora, etc.

The process of answering questions by mining Web data is complicated by the low average quality of individual documents. Many Web documents are poorly written, are barely edited, or simply contain incorrect information. Thus, text extracted from a single document cannot be trusted as the correct answer. Fortunately, this problem can also be alleviated by data redundancy. A single instance of a candidate answer may not provide sufficient justification, but multiple occurrences of the same answer in different documents lends credibility to the proposed answer; voting is a straightforward way to capture this insight. However, majority polling alone is not sufficient to guarantee the correct answer—only the most popular one; although the most popular answer is usually correct, there are notable exceptions. Answer validation and verification is a task that requires further research.
Aranea is a question answering system that implements the above question answering strategy. Using existing Web search engines as a basis, Aranea postprocesses search results using simple statistical techniques that capitalize on the data redundancy of the World Wide Web. Some of these techniques include $n$-gram mining, $n$-gram tiling, and voting (see [1] for more details). In addition to statistical techniques, Aranea also employs a collection of heuristic filters that bring external knowledge (e.g., knowledge about named entities and ontologies) to bear in the question answering process; these filters throw out answer candidates that are obviously wrong. Figure 1 shows a screenshot of the system.

![Aranea Screenshot](image)

**Figure 1: Screenshot of the Aranea system**

**Impact:** Often, data redundancy can serve as a surrogate for language understanding. Without the luxury of massive amounts of data, a question answering system may be forced to extract answers from passages in which they are not obviously stated, e.g., passage (2) in the Lincoln example above. In these cases, sophisticated natural language processing may be required to relate the answer to the question, e.g., recognizing syntactic alternations, resolving anaphora, making commonsense inferences, etc.

Our research with redundancy-based question answering on the Web has convinced us that, at least for some tasks, simple non-linguistically motivated techniques perform as well as their linguistically sophisticated counterparts. Therefore, we believe that a thorough investigation of such techniques is a worthwhile endeavor.

**Future Work:** Although redundancy-based techniques have proven to be effective for extracting answers to fact-based questions, there are fundamental limitations that no amount of clever trickery can hope to solve. Phenomena such as temporal references and embedded clauses require at least the understanding of syntactic structure. To synthesize sensible answers from multiple, possibly conflicting, sources is beyond the reach of simple redundancy-based techniques. Nevertheless, we believe that simple, reliable data mining methods will serve as a baseline for question answering systems.

**Research Support:** This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory.

**References:**
