# The Web as a Resource for Question Answering: Perspectives and Challenges

**Jimmy Lin**

MIT Artificial Intelligence Laboratory
200 Technology Square, Cambridge, MA 02139, USA
jimmylin@ai.mit.edu

**Abstract**

The vast amounts of information readily available on the World Wide Web can be effectively used for question answering in two fundamentally different ways. In the *federated* approach, techniques for handling semistructured data are applied to access Web sources as if they were databases, allowing large classes of common questions to be answered uniformly. In the *distributed* approach, large-scale text-processing techniques are used to extract answers directly from unstructured Web documents. Because the Web is orders of magnitude larger than any human-collected corpus, question answering systems can capitalize on its unparalleled-levels of data redundancy. Analysis of real-world user questions reveals that the federated and distributed approaches complement each other nicely, suggesting a hybrid approach in future question answering systems.

## 1. Introduction

The vast amounts of information readily available on the World Wide Web makes it a very attractive resource for answering simple, fact-based questions such as "Who killed Lincoln?" or "Who discovered x-rays?" Traditionally, question answering has been conducted on relatively small, closed corpora; extending the data source to include information freely available from the Web presents exciting new opportunities and challenges.

I present two fundamentally different approaches of using Web data for question answering, termed *federated* and *distributed*. In the federated approach, portions of the Web are treated as if they were databases, using techniques for managing semistructured data. From this vantage point, question answering translates into a problem of organizing distributed, semistructured information under a natural language interface.

In the distributed approach, Web data is viewed as an enormous collection of unstructured, flat text, with tremendous amounts of data redundancy. Its immense size qualitatively changes the nature of the question answering task, as compared to the same task on closed corpora, e.g., newspaper texts, encyclopedias, etc.

Despite the disparity between the federated and distributed approaches, a crucial observation allows them to complement each other—the types of natural language queries that users ask qualitatively obey Zipf's Law[1] (Zipf, 1935). Analysis of existing question answering test sets reveals that many similar questions occur frequently, e.g., "What is the population of *x*?" These questions translate naturally into database queries, and can be cleanly handled by federation techniques. In addition to large classes of commonly occurring questions, there are also large numbers of unique questions, e.g., "What format was VHS's main competition?" The distributed approach provides a general purpose solution for handling such questions using Web data. Because the strengths of each approach lies in different types of questions, both the federated and distributed approaches can be effectively integrated into a single question answering system.

## 2. The Federated Approach

Although the Web consists largely of unorganized pages, pockets of structured knowledge exist as valuable resources for question answering. For example, the CIA World Factbook provides political, geographic, and economic information about every country in the world; Biography.com contains profiles of over twenty-five thousand famous (and not-so-famous) people; the Internet Movie Database stores entries for hundreds of thousands of movies, including information about their cast, production staff, etc.

To effectively use these existing resources for question answering, the plethora of knowledge sources must be integrated, or federated, under a common interface or query language. Database concepts and techniques provide the tools to accomplish just that. In fact, since many of these sources are part of the "deep" or "invisible" Web, they are inaccessible to search engines and can only be modeled as "virtual" databases. In the spirit of natural language interfaces to relational databases (Androutsopoulos et al., 1995) dating back to the sixties and seventies (Green et al., 1961; Woods et al., 1972; Hendrix, 1977), portions of the Web could be viewed as a semistructured database, serving as the foundation for question answering.

Many existing systems, e.g., ARANEUS (Atzeni et al., 1997), ARIADNE (Knoblock et al., 1999), Information Manifold (Kirk et al., 1995), LORE (McHugh et al., 1997), TSIMMIS (Hammer et al., 1997), just to name a few, have attempted to unify heterogenous Web sources under a common interface. Unfortunately, queries to such systems must be formulated in SQL, Datalog, or some similarly formal language, which render them inaccessible to the average user. Because the focus of research in semistructured data has been on issues such as the modeling of heterogenous

[1] This is a slight abuse of terminology; the question distribution does not exactly obey the reciprocal rank law. I am primarily referring to its general shape.

knowledge sources, the expressiveness of the query language, and implementation issues arising from the unreliable nature of the Web,[2] little work has been done on natural language querying capabilities.

The rich body of research in the management of semistructured data can be leveraged for natural language question answering by writing *schemas* that translate natural language queries, e.g., "What is the population of Taiwan?" into lower-level database queries, e.g.,

```
SELECT population FROM
CIA.countries
WHERE country = 'Taiwan'
```

This is basically the *annotations* technology proposed by Katz (Katz, 1988). Naturally, schemas should be composed at the semantic level in order to handle the complexities of natural language, e.g., structural alternations, `IS-A` and other lexical relations, etc.

### 2.1. Quantifying Performance

To quantify the effectiveness of database federation for question answering, I evaluated the performance of a hypothetical system that utilizes semistructured database techniques against test questions drawn from the TREC-9[3] (Voorhees and Tice, 2000) and TREC-2001 (Voorhees, 2001) QA Track.

After choosing ten knowledge sources widely available on the World Wide Web *a priori*, I manually determined and verified that those sources, taken together, were capable of answering $27\%$ of TREC-9 and $47\%$ of TREC-2001 questions. Table 1 shows a detailed breakdown of the number of questions each site could answer. Furthermore, I confirmed that there were no other large classes of questions that could have been answered by a knowledge source not on the list of ten considered.

For the purposes of this experiment, I assumed that knowledge in the ten chosen sources have been integrated under a common interface (for concreteness, I assumed a SQL-like query language). A question, e.g., "What is the Ohio state bird," would be translated into a database query at the logical form level by a schema, e.g.,

**state-bird**$(x) \rightarrow$
`SELECT bird FROM 50states.states`
`WHERE state =` $x$

The ability to parse question accurately into semantic representations is within limits of current state-of-the-art systems. By treebanking questions, parse accuracy can be boosted to above 95% (Hermjakob, 2001). Thus, schemas like the example above can be practically used to connect natural language queries to a semistructured database.

We can quantify the amount of knowledge engineering necessary to achieve the abovementioned level of performance in this hypothetical system by calculating the number of schemas required, since that number is roughly proportional to the amount of manual labor required (Figure 2).

**definition**$(x) \rightarrow$
`SELECT def FROM dictionary.defs`
`WHERE word =` $x$
**32** Questions from TREC-9, e.g.,
(228) What is platinum?
(300) What is leukemia?
**129** Questions from TREC-2001, e.g.,
(980) What is amoxicillin?
(994) What is neurology?

**capital**$(x) \rightarrow$
`SELECT population FROM CIA.countries`
`WHERE country =` $x$
**6** Questions from TREC-9, e.g.,
(205) What is the population of the Bahamas?
(365) What is the population of Mozambique?
**4** Questions from TREC-2001, e.g.,
(1120) What is the population of Nigeria?
(1238) What is the population of Australia?

Figure 1: Sample schemas

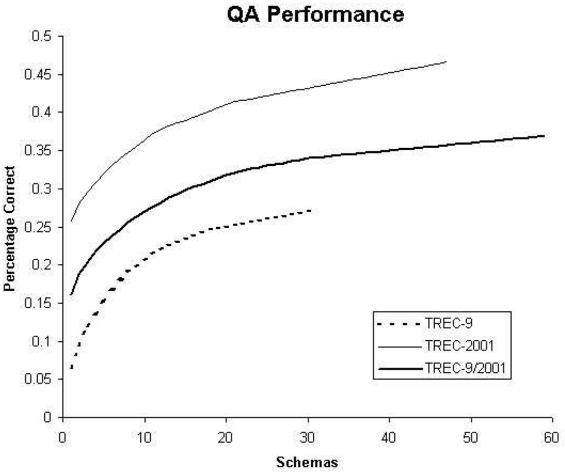

Figure 2: Question Answering performance graphed against number of schemas in a federated-database approach

The graph reveals that a large number of questions can be handled by a relatively small number of schemas. Two sample schemas are shown in Figure 1. The TREC test sets, reflective of real-world user queries, display an interesting characteristic: a large number of questions share the same form, e.g., "What is $x$", "Who is $x$", etc.[4] Database federation is very effective at handling these large classes of similar questions.

The hypothetical system discussed above is by no means a distant fantasy; in fact, an example of such systems has existed for a while. START[5] (Katz, 1988; Katz, 1997) has been answering user questions on the World Wide Web since 1993, and uses, among other techniques, the federated

[2] For a survey of database techniques for the Web, see (Florescu et al., 1998).

[3] the first five hundred questions, not including the variants

[4] Dictionary.com and biography.com, respectively, provide excellent answers to these questions.

[5] `http://www.ai.mit.edu/projects/infolab`

| Test Set | BIO | CIA | DIC | INF | POT | 50S | Other | Total | Score |
|---|---|---|---|---|---|---|---|---|---|
| TREC-9 (500 questions) | 28 | 24 | 55 | 7 | 3 | 11 | 9 | 137 | .27 |
| TREC-2001 (500 questions) | 11 | 26 | 141 | 14 | 12 | 15 | 15 | 234 | .47 |
| Total (1000 questions) | 39 | 50 | 196 | 21 | 15 | 26 | 24 | 371 | .37 |

Table 1: Breakdown, by knowledge source, of correctly answered TREC questions. BIO = biography.com, CIA = CIA World Factbook, DIC = dictionary.com, INF = Infoplease.com, POT = Presidents of the United States (POTUS), 50S = 50states.com

data integration approach. Omnibase (Katz et al., 2001) is the database engine that serves as START's gateway to Web data sources.

## 3. The Distributed Approach

Despite the effectiveness of database federation, structured knowledge accounts for an exceedingly small fraction of the Web. Most of the Web is still, and will most likely remain, unstructured, textual documents. The distributed approach utilizes techniques for dealing with vast amounts of text, which grew out of question answering systems that operated on unstructured corpora, e.g., newspaper archives, encyclopedias, etc. The dominant approach to extracting answers from a closed corpus, driven primarily by information retrieval and information extraction technology, is basically a two-step process:

1. Reduce the corpus to a smaller set of relevant documents (or segments therefrom) using, for example, passage retrieval techniques.

2. Attempt to "pinpoint" the exact location of the answer. One successful approach is to search for an entity whose semantic type matches the type extracted from the question. For example, a "who" question would trigger a search for person names, and a "when" question would trigger a search for dates.

In moving from corpus-based to Web-based question answering, there is an important questions to consider. Does the immense difference in collection size qualitatively change the task of question answering, thereby necessitating the development of new techniques?

In short, yes! As a text collection, the Web is staggering in size, to the point where researchers cannot even agree on a methodology for measuring its size. Although estimates vary, consensus places the size of text on the Web in the tens of terabytes range (Lyman and Varian, 2000; Bright-Planet Corporation, 2001). Google, the largest of the Web search engines, indexes a staggering 2 billion Web documents,[6] which is still only a fraction of its entirety. The Web's size dwarfs any human-collected corpus by several orders of magnitude. An important implication of this size is the amount of data redundancy inherent in the text collection; potentially, each item of information has been stated in a variety of ways, in different documents. However, this is counterbalanced by the poor quality of individual documents. Due to these unique characteristics, question answering techniques that were effective on closed corpora cannot be blindly applied to Web data.

The tremendous amounts of information on the World Wide Web would be useless without an effective method of data access. However, providing the basic infrastructure for indexing and retrieving text at such scales is a tremendous engineering task. Fortunately, such services already exist, in the form of search engines. Using them as a (relatively primitive) IR backend, question answering systems can capitalize on data redundancy in two major ways: as a surrogate for sophisticated natural language techniques and as a method for overcoming poor document quality.

### 3.1. Redundancy as Surrogate for Understanding

In the TREC evaluations, Breck and his colleagues (Breck et al., 2001) noted a correlation between the number of times an answer appeared in the test corpus and the average performance of question answering systems on that particular question. This result verifies intuition: the more frequently an answer appears, the easier it is to find it. Extending this to the World Wide Web, it is fairly obvious to see how massive amounts of data can be effectively used for question answering.

Consider the question "Who killed Lincoln?" Here are two possible answers:

> (1) John Wilkes Booth killed 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 sentence (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. In these cases, simple techniques, e.g., keyword-based passage retrieval (Harabagiu et al., 2000a; Harabagiu et al., 2001; Clarke et al., 2001a; Clarke et al., 2001b; Hovy et al., 2001), serve as an adequate foundation for achieving state-of-the-art performance. 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.

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 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.

Surprisingly, the World Wide Web is so big that simple pattern matching techniques can replace the need to un-

[6] as of early March, 2002

derstand both the structure and meaning of language. The answer to a question could be extracted by searching directly for an anticipated answer form, e.g., in the above example, by searching for the string "killed Lincoln" and extracting words occurring to the left (Kwok et al., 2001; Brill et al., 2001). 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 (Clarke et al., 2001a). Data redundancy enables a simple trick to overcome many troublesome issues in natural language processing, e.g., alternations, anaphora, etc. In fact, simple pattern matching techniques have already been applied very successfully to the TREC corpus (Soubbotin and Soubbotin, 2001); applying the same tricks to Web data promises further boost in performance.

### 3.2. Redundancy and Answer Quality

The process of answering questions using Web data is complicated by the low average quality of individual documents. Due to the low barrier of entry in Web publishing, many documents are poorly written, barely edited, or simply contain incorrect information. As a result, text extracted from a single document cannot be trusted as the correct answer. This problem can also be alleviated through 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 (Kwok et al., 2001; Brill et al., 2001; Buchholz, 2001; Clarke et al., 2001a) is a straightforward way to use data redundancy to verify proposed answers. In contrast with approaches that use linguistically sophisticated techniques such as abductive proofs (Harabagiu et al., 2000a; Harabagiu et al., 2000b), voting requires no external domain-specific knowledge, and is easy to implement.

### 3.3. Implemented Systems

The effectiveness of Web-driven question answering techniques can be seen at the TREC-2001 QA Track (Voorhees, 2001), which evaluated systems from thirty-six different teams around the world. The Microsoft Research entry relied exclusively on Web data retrieved through a standard Web search engine, utilizing simple techniques such as pattern matching, *n*-gram generation, and counting to generate and confirm answers (Brill et al., 2001). The MSR system was among the top performers in the evaluation, which demonstrated that with the entire Web at the disposal of a question answering system, a relatively simple system approaches the performance of state-of-the-art knowledge-intensive systems, e.g., (Harabagiu et al., 2000a; Hovy et al., 2000; Harabagiu et al., 2001; Hovy et al., 2001). Another one of the top systems (Clarke et al., 2001b), by the University of Waterloo, used the Web as a secondary corpus to verify answers directly retrieved from the TREC corpus. Web reinforcement of answers boosted performance of the system by 25% over the same baseline system without Web-support answers.

In addition to entries at TREC-2001, many Web-based question answering programs have been developed. MULDER (Kwok et al., 2001), AnswerBus, and NSIR (Radev et al., 2002) are examples of systems that operate by postprocessing results from standard Web search engines. Ionaut (Abney et al., 2000) uses information extraction techniques to pinpoint answers in its local collection of Web pages. The FAQ Finder System (Burke et al., 1997) takes advantage of files containing frequently-asked questions widely available on the Web. Because these systems, with the exception of NSIR, have not been evaluated against a standard test set (e.g., TREC), it is difficult to compare their overall performance.

## 4. Discussion

The federated and distributed approaches to Web-based question answering should be viewed as complementary, not competing, strategies that focus on different aspects of the World Wide Web. Treating the Web as a semistructured database is an excellent strategy for handling large, predictable classes of questions with parametric variations. The distributed approach is a much more general-purpose solution to question answering, capable of achieving broad coverage without labor-intensive knowledge engineering. If we assume that query distribution in the question answering task obeys Zipf's Law,[7] then the federated approach is well-suited for the head of the Zipf curve, and the distributed approach provides capabilities for handling its broad tail.

### 4.1. Challenges and Issues with Federation

Naturally, the biggest challenge to the integration of multiple knowledge sources from the Web is the lack of explicit and uniform database schemas. Mechanisms must be created to mediate the interaction between the query language and individual knowledge sources. This is usually accomplished through site-specific *wrappers* that translate local data into a form digestible by the integration system. Often, wrappers are as simple as pattern matching scripts, but they can be, nevertheless, time-consuming and laborious to construct, especially for large knowledge sources. The amount of manual labor required for database federation is one of its critical shortcomings. However, there are promising solutions to this problem. To start, a simple, well-designed authoring tool, e.g., (Adelberg, 1998; Sahuguet and Azavant, 1999), can drastically reduce the amount of time required to integrate a knowledge source. In addition, machine learning techniques can be applied to automate the wrapper generation process (Kushmerick et al., 1997; Hsu and Chang, 1999; Muslea et al., 1999). The most promising solution to this problem is Semantic Web (Berners-Lee, 1999) research, which seeks to augment ordinary Web documents with semantic annotations and other metadata. If this dream is ever realized, integration of multiple knowledge sources could be accomplished effortlessly.

Another issue with treating the Web as a semistructured database is the problem of limited coverage. Since structured knowledge exists on the Web in the form of domain-

[7] which appears so, at least for TREC-style questions

specific sites, it is difficult to achieve broad coverage using the federated approach exclusively without significant amounts of manual labor. To make matters worse, since the Web is still predominately flat text files, structure cannot be assigned to (or derived from) most documents, rendering database techniques useless. Fortunately, analysis of realistic natural language queries reveals that users often ask the same types of questions over and over again; in fact, relatively few schemas are sufficient to achieve moderate amounts of coverage (Figure 2). Naturally, the federated approach does have its limits, due to diminishing returns as the number of schemas increase.

Because each federated knowledge source is domain-specific and (generally) well-structured, there is greater certainty that coverage within a particular domain is complete. As a simple example, a corpus might contain the answer to "How long is the coastline of Canada," but there is no guarantee that the same information exists for all two hundred countries in the world. In fact, it is doubtful that any corpus would describe landlocked countries as having coastlines zero kilometers long.[8] Because structured information sources usually represent significant effort in knowledge engineering, coverage for parametric variations within question classes is generally quite high, e.g., having coastline lengths for *all* countries, having population figures for *all* states, etc. Furthermore, federated queries can be delegated to authoritative and trustworthy sources to ensure the quality of the answers, e.g., census bureaus for population figures, geographic/geological surveys for coastlines, etc.

Federation of Web sources allows the creation of *virtual* databases (Hull and Zhou, 1996; Hull, 1996), where the actual knowledge is distributed around the Web and retrieved at query-time. In a sense, a federated system acts like a knowledge broker, much like a librarian at the reference desk of a large research library. Although such a methodology introduces challenges stemming from network unreliability, it provides the dual advantage of simpler, distributed maintenance and seamless, centralized integration. Furthermore, federation allows up-to-date access to timely information, e.g., headlines, stock prices, weather, etc.

### 4.2. Challenges and Issues with Distribution

A major advantage of the distributed approach to Web-based question answering is the generality of the solution. Coupled with external resources, e.g., a typology of question types, a large variety of questions can be answered in a relatively uniform framework. In a sense, this line of research is merely an extension of recent trends in empirical natural language processing on very large corpora. It has been shown that for some natural language tasks (e.g., word-context disambiguation), performance can be greatly improved by simply acquiring and using more training data (Banko and Brill, 2001); an open research question is to what extent question answering can benefit from simply having more data, and how far simple techniques can be pushed.

Fueling the work in Web-driven question answering are parallel advances in the immense engineering tasks of crawling and indexing the entire Web, e.g., (Brin and Page, 1998). Fortunately, suitable infrastructure already exists, in the form of search engines, for retrieving large amounts of Web data for further postprocessing. Although they may prove to be insufficient for question answering purposes, a Web search engine like Google can nevertheless serve as an immediate testbed for experimental purposes.

Despite its generality, the distributed approach is not suitable for certain classes of questions,[9] e.g., definition questions ("What is leukemia?"). In total, these questions constituted roughly a quarter of the TREC-2001 test set. With the distributed approach, it is very difficult to control answer quality and respond appropriately to different types of users. Prager (Prager et al., 2001) has noted that without an accurate user model, it is difficult to determine what constitutes a "good answer." More generally, Lehnert (Lenhert, 1981) has argued that without an accurate model of both the questioner and the answerer, systems are liable to return pragmatically-incorrect answers. With definition questions, a hypernym, a dictionary definition, a general encyclopedia entry, or a domain-specific article might all be appropriate answers, depending on the user. The distributed approach might plausibly provide a hypernym or dictionary-definition answer, but more detailed answers are beyond the scope of current natural language technology, for they require the ability to integrate information from multiple documents into a coherent whole—basically, multi-document summarization, e.g., (Radev and McKeown, 1998). The federated approach offers a nice solution to the answer quality problem. Instead of using unstructured flat corpora, why not use rich knowledge sources that already exist? The choice of knowledge source from which to answer a specific question would be decided based on user modeling ("content specification" in Lehnert's terms), e.g., a general encyclopedia article for a high-school student doing research for a school report, a simple dictionary definition for the casual user, etc. The same technique can be applied to handle other large generic question classes, e.g., "Where is $x$" and "Who is $x$".

The amount of data redundancy available on the Web allows sophisticated natural language processing techniques to be replaced by simple pattern matching techniques. Nevertheless, these patterns still need to be computed from the questions, e.g., a system must generate the pattern "Belize is located in . . . " from "Where is Belize located?" The method could be as simple as trying all plausible permutations of the query (Brill et al., 2001), but more principled methods should be sought after. For this task, machine learning will likely play an important role. In fact, techniques for automatically acquiring search engine reformulations (Agichtein and Gravano, 2001; Agichtein et al., 2001; Radev et al., 2001) and lexical inference rules (Lin and Pantel, 2001) have already been explored, as well as tricks to get the most out of existing search engines by simple query expansion (Magnini and Prevete, 2000).

Although voting is a simple technique to verify answers,

[8] The alternative to answering such questions, inferring "no coastline" from "landlocked," requires domain-specifi c knowledge and is also impractical for open-domain question answering.

[9] especially in the 50-byte TREC format

the most frequently occurring answer is sometimes not the correct one. A memorable example is the question "Who invented the paper clip," to which a Web-based question answering system answered "Trent Lott," who claimed that the invention was his in response to Al Gore's claim that he invented the Internet. To address these issues, several solutions are being explored. In one approach, Web data is exploited more conservatively; Clarke and his colleagues (Clarke et al., 2001b) collect Web documents into a secondary corpus to boost results obtained from a primary, more authoritative corpus. Another solution is to score Web pages based on its quality and authority (Page et al., 1998; Amento et al., 2000; Zhu and Gauch, 2000).

Even with the amount of Web data, there are fundamental problems that require true natural language processing to solve. Pattern matching is simple but dangerous, because the technique is insensitive to linguistic constructions such as constituent boundaries and embedded clauses. Redundancy helps, but is not a panacea. A Web-driven question answering system once returned "Steven Spielberg" to the question "Who is the prime minister of Israel," from the perfectly valid sentence (in the context of jokes) "George Bush thinks Steven Spielberg is the Prime Minister of Israel." Temporal questions also pose difficulty to the linguistically naive system, e.g., "Who was the president of the United States in 1992?" Checking the timestamp on the document alone is insufficient, because time could be relative to the date on which the document was authored, e.g., "Three years ago, Bill Clinton was the president of the United States" (in a document dated 2002). Obviously, Web-driven approaches could benefit from natural language processing techniques to overcome many of these problems, e.g., the application of syntactic relations to boost accuracy (Litkowski, 1999; Attardi et al., 2001; Buchholz, 2001; Lin, 2001; Litkowski, 2001), abductive reasoning to justify answers (Harabagiu et al., 2000a; Harabagiu et al., 2000b), and coreference resolution (Morton, 1999).

In the context of TREC evaluation, an issue with which Web-driven QA systems have had to contend is the problem of finding support for answers derived from the Web. Since TREC requires a supporting document drawn from its corpus, systems have had to "project" Web-derived answers back onto the TREC corpus (Brill et al., 2001; Hovy et al., 2001). One might argue that the only purpose of this artificial requirement is to facilitate the judgment process, but this issue of "how to answer a question if you already know the answer" is not as ridiculous as it might sound. For starters, the projection process serves to validate the answer. But there are other realistic scenarios: for example, a content supplier on the Web would want to answer user queries with knowledge from its own site (for obvious business reasons); with answer projection, this could be accomplished while simultaneously taking advantage of the entire Web as a knowledge resource.

## 5. The Future

Naturally, a question answering system could benefit from adopting a combination of both the federated and distributed approaches. Because fact-based questions appear to obey Zipf's Law and the two approaches show their strengths at different ends of the Zipf Curve, the philosophically different techniques complement each other nicely.

Although information retrieval components still serve as the foundation of most question answering systems today, one can already observe a trend in the use of external knowledge resources, the natural extension of which is database federation. For example, Webclopedia (Hovy et al., 2001) uses WordNet (Miller, 1995) to assist in answering definition questions. IBM's statistical question answering system uses an external encyclopedia (Ittycheriah et al., 2000; Ittycheriah et al., 2001) for query expansion. I believe that the federated approach will continue to gain popularity, such that future systems will be a hybrid composition of both the federated and distributed approach. As we turn to more difficult forms of question answering, as outlined in roadmaps (Carbonell et al., 2000; Burger et al., 2000), it would be very interesting to see the role that each approach will play.

Although the availability of extremely large corpora presents exciting opportunities for question answering, this optimism is tempered by challenges yet to be resolved both with the federated and distributed approaches. Although more research will be required, the impact of the World Wide Web on question answering and other applications will be no less than revolutionary.

## 6. Acknowledgements

I'd like to thank Boris Katz, Greg Marton, and Vineet Sinha for their helpful comments on earlier drafts.

## 7. References

Steven Abney, Michael Collins, and Amit Singhal. 2000. Answer extraction. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*.

Brad Adelberg. 1998. NoDoSE—a tool for semi-automatically extracting structured and semistructured data from text documents. *SIGMOD Record*, 27:283–294.

Eugene Agichtein and Luis Gravano. 2001. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries (DL'00)*.

Eugene Agichtein, Steve Lawrence, and Luis Gravano. 2001. Learning search engine specific query transformations for question answering. In *Proceedings of the 10th International World-Wide Web Conference (WWW10)*.

Brian Amento, Loren Terveen, and Will Hill. 2000. Does authority mean quality? predicting expert quality ratings of Web documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000)*.

Ion Androutsopoulos, Graeme D. Ritchie, and Peter Thanisch. 1995. Natural language interfaces to databases—an introduction. *Natural Language Engineering*, 1(1):29–81.

Giuseppe Attardi, Antonio Cisternino, Francesco Formica, Maria Simi, Alessandro Tommasi, and Cesare Zavattari. 2001. PIQASso: Pisa question answering system. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Paolo Atzeni, Giansalvatore Mecca, and Paolo Merialdo. 1997. Semistructured and structured data in the Web: Going back and forth. In *Proceedings of the Workshop on Management of Semistructured Data at PODS/SIGMOD'97*.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*.

Tim Berners-Lee. 1999. *Weaving the Web*. Harper, New York.

Eric Breck, Marc Light, Gideon S. Mann, Ellen Riloff, Brianne Brown, Pranav Anand, Mats Rooth, and Michael Thelen. 2001. Looking under the hood: Tools for diagnosing your question answering engine. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01) Workshop on Open-Domain Question Answering*.

BrightPlanet Corporation. 2001. Deep Web white paper.

Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais, and Andrew Ng. 2001. Data-intensive question answering. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. In *The 6th International World Wide Web Conference*.

Sabine Buchholz. 2001. Using grammatical relations, answer frequencies and the World Wide Web for question answering. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. 2000. Issues, tasks, and program structures to roadmap research in Question & Answering (Q&A).

Robin D. Burke, Kristian J. Hammond, Vladimir A. Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg. 1997. Question answering from frequently-asked question files: Experiences with the FAQ finder system. Technical Report TR-97-05, University of Chicago.

Jamie Carbonell, Donna Harman, Eduard Hovy, Steve Maiorano, John Prange, and Karen Sparck-Jones. 2000. Vision statement to guide research in Question & Answering (Q&A) and Text Summarization.

Charles Clarke, Gordon Cormack, and Thomas Lynam. 2001a. Exploiting redundancy in question answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*.

Charles Clarke, Gordon Cormack, Thomas Lynam, C.M. Li, and Greg McLearn. 2001b. Web reinforced question answering (MultiText experiments for TREC 2001). In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Daniela Florescu, Alon Levy, and Alberto Mendelzon. 1998. Database techniques for the World-Wide Web: A survey. *SIGMOD Record*, 27(3):59–74.

Bert Green, Alice Wolf, Carol Chomsky, and Kenneth Laughery. 1961. BASEBALL: An automatic question answerer. In *Proceedings of the Western Joint Computer Conference*.

Joachim Hammer, Hector Garcia-Molina, Junghoo Cho, Rohan Aranha, and Arturo Crespo. 1997. Extracting semistructured information from the Web. In *Proceedings of the Workshop on Management of Semistructured Data at PODS/SIGMOD'97*.

Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. 2000a. FALCON: Boosting knowledge for answer engines. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*.

Sanda Harabagiu, Marius Paşca, and Steven Maiorano. 2000b. Experiments with open-domain textual question answering. In *Proceedings of the 18th Annual International Conference on Computational Linguistics (COLING-2000)*.

Sanda Harabagiu, Dan Moldovan, Marius Paşca, Mihai Surdeanu, Rada Mihalcea, Roxana Gîrju, Vasile Rus, Finley Lăcătuşu, Paul Morărescu, and Răzvan Bunescu. 2001. Answering complex, list, and context questions with LCC's Question-Answering Server. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Gary G. Hendrix. 1977. Human engineering for applied natural language processing. Technical Note 139, SRI International.

Ulf Hermjakob. 2001. Parsing and question classification for question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01) Workshop on Open-Domain Question Answering*.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. 2000. Question answering in Webclopedia. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*.

Eduard Hovy, Ulf Hermjakob, and Chin-Yew Lin. 2001. The use of external knowledge in factoid QA. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Chun-Nan Hsu and Chien-Chi Chang. 1999. Finite-state transducers for semi-structured text mining. In *Proceedings of the IJCAI-99 Workshop on Text Mining: Foundations, Techniques, and Applications*.

Richard Hull and Gang Zhou. 1996. A framework for supporting data integration using the materialized and virtual approaches. In *Proceedings of the ACM SIGMOD Conference on Management of Data*.

Richard Hull. 1996. Managing semantic heterogeneity in databases. In *Proceedings of the ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems (PODS'96)*.

Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi. 2000. IBM's statistical question answering system. In *Proceedings of the 8th Text REtrieval Conference (TREC-9)*.

Abraham Ittycheriah, Martin Franz, and Salim Roukos. 2001. IBM's statistical question answering system—TREC-10. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Boris Katz, Deniz Yuret, and Sue Felshin. 2001. Omnibase: A universal data source interface. In *MIT Artificial Intelligence Laboratory Abstracts*.

Boris Katz. 1988. Using English for indexing and retrieving. In *Proceedings of the 1st RIAO Conference on User-Oriented Content-Based Text and Image Handling (RIAO '88)*.

Boris Katz. 1997. Annotating the World Wide Web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*.

Thomas Kirk, Alon Levy, Yehoshua Sagiv, and Divesh Srivastava. 1995. The Information Manifold. Technical report, AT&T Bell Laboratories.

Craig Knoblock, Steven Minton, Jose Luis Ambite, Naveen Ashish, Ion Muslea, Andrew Philpot, and Sheila Tejada. 1999. The Ariadne approach to Web-based information integration. *International the Journal on Cooperative Information Systems (IJCIS) Special Issue on Intelligent Information Agents: Theory and Applications*, 10(1/2):145–169.

Nickolas Kushmerick, Daniel Weld, and Robert Doorenbos. 1997. Wrapper induction for information extraction. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*.

Cody Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the Web. In *Proceedings of the Tenth International World Wide Web Conference (WWW10)*.

Wendy G. Lenhert. 1981. A computational theory of human question answering. In Aravind K. Joshi, Bonnie L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 145–176. Cambridge University Press, Cambridge, England.

Dekang Lin and Patrick Pantel. 2001. DIRT—discovery of inference rules from text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Jimmy J. Lin. 2001. Indexing and retrieving natural language using ternary expressions. Master's thesis, Massachusetts Institute of Technology.

Kenneth C. Litkowski. 1999. Question-answering using semantic relation triples. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*.

Kenneth C. Litkowski. 2001. CL Research experiments in TREC-10 question answering. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Peter Lyman and Hal R. Varian. 2000. How much information, October.

Bernardo Magnini and Roberto Prevete. 2000. Exploiting lexical expansions and boolean compositions for Web querying. In *Proceedings of the ACL 2000 Workshop on Recent Advances in NLP and IR*.

Jason McHugh, Serge Abiteboul, Roy Goldman, Dallan Quass, and Jennifer Widom. 1997. Lore: A database management system for semistructured data. Technical report, Stanford University Database Group, February.

George Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):49–51.

Thomas S. Morton. 1999. Using coreference in question answering. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*.

Ion Muslea, Steve Minton, and Craig Knoblock. 1999. A hierarchical approach to wrapper induction. In *Proceedings of the 3rd International Conference on Autonomous Agents*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The pagerank citation ranking: Bringing order to the web.

John Prager, Jennifer Chu-Carroll, and Krzysztof Czuba. 2001. Use of WordNet hypernyms for answering what-is questions. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Dragomir Radev and Kathleen McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Dragomir Radev, Hong Qi, Zhiping Zheng, Sasha Blair-Goldensohn, Zhu Zhang, Waiguo Fan, and John Prager. 2001. Mining the Web for answers to natural language questions. In *ACM CIKM 2001: Tenth International Conference on Information and Knowledge Management*.

Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal. 2002. Probabilistic question answering on the web. In *Proceedings of the Eleventh International World Wide Web Conference*.

Arnaud Sahuguet and Fabien Azavant. 1999. Wysi-Wyg Web Wrapper Factory (W4F). In *Proceedings of the Eighth International World Wide Web Conference (WWW8)*.

M.M. Soubbotin and S.M. Soubbotin. 2001. Patterns of potential answer expressions as clues to the right answers. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*.

Ellen M. Voorhees and Dawn M. Tice. 2000. Overview of the TREC-9 question answering track. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*.

Ellen M. Voorhees. 2001. Overview of the TREC 2001 question answering track. In *Proceedings of the 2001 Text REtrieval Conference (TREC 200)*.

W.A. Woods, R.M. Kaplan, and B.N. Webber. 1972. The Lunar Sciences Natural Lanaugage Information System: Final report. Technical Report 2378, BBN.

Xiaolan Zhu and Susan Gauch. 2000. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000)*.

George Kingley Zipf. 1935. *The Psycho-Biology of Language*. MIT Press, Cambridge, Massachusetts.