START, Omnibase, and Beyond

Boris Katz Jimmy J. Lin
MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
{boris,jimmylin}@ai.mit.edu

Abstract

To address the problem of information overload in today's world, we have developed START, a natural language question answering system that provides users with multimedia information access through the use of natural language annotations. In order to harness the potential of knowledge sources on the World Wide Web, we have developed Omnibase, a virtual database that provides uniform access to Web resources. Our ultimate goal is to develop a computer system that acts like a "smart reference librarian," and to a large extent, we have accomplished our goal. This paper describes our current implemented system and discusses future research directions.

1 Introduction

We believe that natural language is the best information access mechanism for humans. We take question answering (QA) to be more than just retrieving relevant text snippets in response to user queries; instead, we envision question answering as a general way of interacting with smart systems using everyday language to access "just the right information."

How can we build systems that provide natural language information access? The intuitive approach would be to take all available information, e.g., all the material in the Library of Congress, the entire World Wide Web, etc., analyze its content, and create a database containing representational structures that capture the "meaning" of the indexed material. A user question would be translated into a "semantic request," and matched against the contents of this database. Regrettably, this rather futuristic scenario is yet sheer fantasy. Unrestricted full-text understanding is beyond the state of the art in natural language processing. Furthermore, not all information is text; images, movies, and sounds can all be valuable sources of knowledge. "Understanding" all these various media would require spectacular breakthroughs in other areas of artificial intelligence, e.g., object recognition, scene analysis, speech transcription, etc. In short, we are still years away from machines capable of distilling "meaning" from various types of documents.

Faced with the limitations of current technology and the insatiable thirst of users for more knowledge, what can we do? Rather than trying to cram all available knowledge in a computer program, we could instead teach the computer *where* and *how* to find the right pieces of knowledge. Such a system would act much like a librarian in the reference section of a library; although she might not be able to answer a question directly, the librarian would nevertheless be helpful because she knows where to find the relevant knowledge. In a sense, we need to give our systems *knowledge about the knowledge*.

2 Natural Language Annotations

How can we create a computer system that acts like a smart reference librarian? A solution is natural language annotations [8, 9, 5], which are machine-parsable sentences and phrases that describe the content of various information segments. They serve as metadata describing the types of questions that a particular piece of knowledge is capable of answering.

To illustrate how this technology works, consider the following paragraph about Olympus Mons:

> The highest point on Mars is Olympus Mons, a huge shield volcano about 26 km (16 miles) high and 600 km (370 mi) across, about the same area as Arizona. Olympus Mons is also the largest known volcano in the solar system.

It may be annotated with the following English sentences and phrases:

> Mars' highest point
> Largest volcano in the solar system.
> Olympus Mons is very tall.

A question answering system would parse these annotations and store the parsed structures (e.g., ternary expressions [4]) with pointers back to the original information segment. To answer a question, the user query

would be compared against the annotations stored in the knowledge base. Because this match would occur at the level of syntactic structures, linguistically sophisticated machinery such as synonymy/hyponymy, ontologies, and structural transformation rules could be brought to bear on the matching process. Linguistic techniques would allow a system to achieve capabilities beyond simple keyword matching, for example, handling complex syntactic alternation involving verb arguments. If a match were found between ternary expressions derived from annotations and those derived from the query, the segment corresponding to the annotations would be returned to the user as the answer. For example, the annotations above would allow a question answering system to answer the following questions (see Figure 1 for an example):

What is the highest point on Mars?
Do you know anything about Olympus Mons?
How tall is Olympus Mons?
Tell me how big Olympus Mons is.
What is the biggest volcano in the solar system?

An important feature of the annotation concept is that any information segment can be annotated: not only text, but also images, multimedia, and even procedures!

We have implemented the above technology in START[1] [4, 5], the first natural language question answering system available on the World Wide Web. Since it came on-line in December, 1993, START has engaged in millions of exchanges with hundreds of thousands of users all over the world, supplying them with useful knowledge.

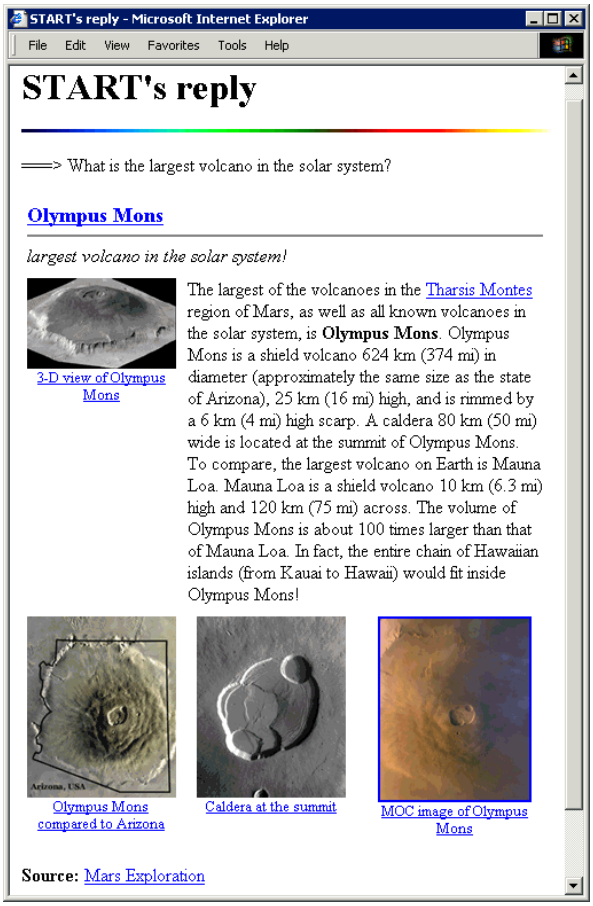

Figure 1: START, our question answering system, responding to the question "What is the largest volcano in the solar system?" with an information segment containing both text and images.

3 The Web as a Database

Although the Web is predominately comprised of static documents, pockets of structured knowledge exist, capable of providing answers to a large number of questions. For example, the CIA World Factbook provides political, geographic, and economic information about every country in the world; Biography.com contains profiles for over twenty-five thousand famous (and not-so-famous) people; the Internet Movie Database contains entries for hundreds of thousands of movies, including information about their cast, production staff, etc.

A question answering system could access this wealth of knowledge uniformly if on-line sources could be treated like a database. We built Omnibase [6], a "virtual" database that does just this: it serves as a structured query interface to heterogeneous data on the World Wide Web. Omnibase accomplishes data integration through the use of site-specific wrapper scripts that mediate between our uniform query language and that of each individual data source.

It is of course impossible to impose any uniform schema on the entire Web. To address this challenge, Omnibase adopts a stylized data relational model which we call the "object–property–value" model. Under this framework, data sources contain *objects* which have *properties*, and questions are translated into requests for the *value* of these properties.

Natural language commonly employs an 'of' relation or a possessive to express the relationship between an object and its property, e.g., "the director of La Strada" or "La Strada's director." Table 1 shows, however, that there are many alternative ways to ask for the value of a property of an object.

Clearly, many other possible types of queries do not fall into our object–property–value model. However, our experiments reveal that in practice questions of the object–property–value type occur quite frequently. For example, just ten Web sources fashioned in the object–property–value manner turned out to be sufficient for handling 27% of TREC-9 and 47% of TREC-2001 questions from the QA track [12].

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

Question	Object	Property	Value
Who wrote the music for Star Wars?	Star Wars	composer	John Williams
Who invented dynamite?	dynamite	inventor	Alfred Nobel
How big is Costa Rica?	Costa Rica	area	51,100 sq. km.
How many people live in Kiribati?	Kiribati	population	94,149
What languages are spoken in Guernsey?	Guernsey	languages	English, French
Show me paintings by Monet.	Monet	works	[images]

Table 1: Some sample questions that can be handled by an object–property–value model of Web data.

Figure 2: START answering the question "Who wrote the screenplay for Good Will Hunting?" by extracting information from the Internet Movie Database and generating an appropriate response.

To actually answer user questions, the gap between natural language questions and structured Omnibase queries must be bridged. Natural language annotations serve as the enabling technology that allows the integration of START and Omnibase. Since annotations can describe arbitrary fragments of knowledge, there is no reason why they can't be employed to describe Omnibase procedures. In fact, annotations can be parameterized, i.e., they can contain symbols representative of an entire class of objects. For example, the annotation "a person wrote the screenplay for `imdb-movie`" can be attached to an Omnibase procedure that retrieves the writers for various movies from the Internet Movie Database (IMDb). The symbol `imdb-movie` serves as a placeholder for any one of the hundreds of thousands of movies that IMDb contains information about; when the annotation matches the user question, the actual movie name is instantiated and passed along to the Omnibase query. After Omnibase fetches the correct answer, START performs additional postprocessing, e.g., natural language generation, to present the answer (see Figure 2).

In September, 1998, we integrated Omnibase with START, giving the combined system access to semistructured Web sources in a uniform framework. Currently, our system can answer millions of natural language questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more.

4 Beyond START and Omnibase

Because START performs sophisticated syntactic and semantic processing of questions to pinpoint the exact information need of a user, questions can be answered with remarkable precision. In the period from January, 2001 to March, 2002, START and Omnibase answered over 326 thousand queries from users all over the world. Of those, 67% were answered successfully by our system (59% of the questions answered were handled by Omnibase).

Despite the effectiveness of START and Omnibase in solving user information needs, expanding their knowledge coverage is a time-consuming process. Annotations must be manually written to describe various knowledge segments; integrating Web sources under Omnibase requires site-specific wrapper scripts. Consequently, only trained individuals can add knowledge to START, but we are currently pursuing several different research paths that attempt to overcome the limitations of our technology, two of which we describe in this paper. First, by providing a collaborative framework in which ordinary users can annotate data, we can distribute and parallelize the annotations process; this work complements current Semantic Web research. Second, we are working on more robust data integration techniques that can provide faster and simpler ways to access heterogeneous data sources.

5 Distributed Annotations

A salient feature of annotation technology is its simplicity. Because information segments can be described in everyday English, ordinary users with no technical skills qualify as annotators. As a result, the knowledge engineering bottleneck could be overcome by employing a large number of annotators working in parallel to build a knowledge base. We believe that the Web provides a low-cost mechanism for accomplishing this: by offering a collaborative framework in which users could annotate knowledge, we can engage millions of people all over the world to populate a common knowledge base. In effect, we can distribute the knowledge engineering task across a large user base by allowing each user to teach the system new items of knowledge. We proposed this mechanism

Figure 3: A system that "conceptually segments" knowledge fragments on a Web page.

for gathering annotations from the Web in 1997 [5], and recently we have begun to implement the idea [7].

The idea of distributed annotations is complementary to current research in the Semantic Web [2], whose vision is to convert Web information into a more machine-friendly form through the use of metadata. Because the Semantic Web is ultimately produced by humans and meant primarily for consumption by humans, we believe that a large portion of Semantic Web metadata should be written in natural language itself. Although the general framework of the Semantic Web includes provisions for natural language technology, the actual deployment of such technology remains largely unexplored. In an effort to exploit the synergistic opportunities between the Semantic Web and natural language techniques, we are developing mechanisms to seamlessly weave natural language technology into the Resource Description Framework, the foundation of the Semantic Web.

6 Robust Data Integration

The execution of an Omnibase query involves the application of site-specific scripts that fetch the relevant Web page for an object and extract the value for a particular property. Although most of these scripts are relatively simple (the majority of them consist of simple regular expressions), the proliferation of scripts presents a complex maintenance challenge. This problem is exacerbated by occasional changes in the layout of Web sources, which necessitates the rewriting of all relevant scripts.

We are currently investigating various solutions to these problems. The initial and most obvious solution is to facilitate the data integration process through well-designed authoring tools (e.g., [1, 14]). By automatically preprocessing Web pages to bring potentially relevant sections to the attention of the data integrator, an authoring tool could drastically reduce the amount of time required to add a knowledge source. Furthermore, machine learning techniques can also be applied to automate the wrapper generation process, e.g., [11, 3, 13]. By providing the system with known examples of properties and values, scripts can be induced automatically. This line of research has the added advantage of remaining resilient to site-layout changes; if script inconsistencies are detected, the system would merely retrain itself on the new set of Web pages.

A more general-purpose solution to the data integration problem is to develop systems that are more "aware" of the content and semantics of the page. For example, the computer should contain heuristics regarding ways in which humans typically organize and display information, e.g., a bold heading and accompanying text frequently imply a coherent knowledge segment, a table is often used to organize properties and values, etc. In short, we want a system that can "conceptually segment" a Web page into coherent components automatically. Such a system (see Figure 3) would assist greatly in the data integration process, e.g., by presenting humans with better guesses about page layout in the context of an authoring tool, by simplifying the machine learning task, etc. In fact, we have already experimented with such tools; LaMeTH [10] is a system that attempts to recognize the conceptual structure of a Web page, not

merely its HTML encoding. It provides a scripting interface that allows humans to describe content in terms of layout elements, e.g., paragraphs, lists, and tables, instead of HTML code. By developing "content-aware" systems, we hope to alleviate the problem of data integration.

7 Conclusion

START and Omnibase are complementary components of a question answering system that addresses users' information access needs. By providing a uniform natural language interface to heterogeneous knowledge on the World Wide Web, we can supply users with "just the right information." We believe that structured access to heterogeneous on-line data sources should be a key component of any future natural language question answering system.

Our systems have proven to be a success by a variety of measures. Nevertheless, many directions for future research remain unexplored, some of which we have discussed in this paper. We remain optimistic about the future prospects of natural language information access systems.

8 Acknowledgements

This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory. Additional funding is provided by the Oxygen Project.

References

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

[2] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

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

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

[5] Boris Katz. 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)*, 1997.

[6] Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimmy Lin, Gregory Marton, Alton Jerome McFarland, and Baris Temelkuran. Omnibase: Uniform access to heterogeneous data as a component of a natural language question answering system. In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*, 2002.

[7] Boris Katz, Jimmy Lin, and Sue Felshin. Gathering knowledge for a question answering system from heterogeneous information sources. In *Proceedings of the ACL 2000 Workshop on Human Language Technology and Knowledge Management*, 2001.

[8] Boris Katz and Patrick Winston. Method and apparatus for generating and utilizing annotations to facilitate computer text retrieval, United States Patent No. 5,309,359, 1994.

[9] Boris Katz and Patrick Winston. Method and apparatus for utilizing annotations to facilitate computer retrieval of database material, United States Patent No. 5,404,295, 1995.

[10] Boris Katz, Deniz Yuret, Jimmy Lin, Sue Felshin, Rebecca Schulman, Adnan Ilik, Ali Ibrahim, and Philip Osafo-Kwaako. Integrating large lexicons and Web resources into a natural language query systen. In *Proceedings of the International Conference on Multimedia Computing and Systems (IEEE ICMCS '99)*, 1999.

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

[12] Jimmy J. Lin. The Web as a resource for question answering: Perspectives and challenges. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002.

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

[14] Arnaud Sahuguet and Fabien Azavant. WysiWyg Web Wrapper Factory (W4F). In *Proceedings of the Eighth International World Wide Web Conference (WWW8)*, 1999.