Organizing and Accessing a Comprehensive Knowledge Base Using the World Wide Web

Boris Katz and Jimmy J. Lin
Computer Science and 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 high-precision information access through the use of natural language annotations. To address the difficulty of accessing large amounts of heterogeneous structured and semistructured data, we have developed Omnibase, which assists* START *by integrating Web databases into a single, uniformly structured "virtual database." To address the sheer amount of unstructured information available electronically, we have developed techniques for distilling large amounts of free text into relations that capture the salient aspects of the text. The combination of natural language annotation technology, object–property–value data model, and relation extraction technology allows us to rapidly develop and deploy smart applications for knowledge intensive domains. Our ultimate goal is to develop a computer system that acts like a "smart reference librarian," providing users with "just the right information" in response to questions posed in natural language.*

1. INTRODUCTION

The vast amounts of text, images, and multimedia freely available on the World Wide Web and in other electronic formats offer a rich resource for next generation knowledge application. One such application is information access—providing knowledge to humans in an intuitive manner. We believe that natural language serves as the best knowledge access mechanism for humans. It is intuitive, easy to use, rapidly deployable, and requires no specialized training. A step in that direction is question answering (QA), where a computer responds directly to natural language questions posed by the user. When asked "What country in Africa has the largest population," a computer should be able to respond with something like "Nigeria, with a population of 126 million, is the most populous African nation." Similarly, the computer should return digital images of Monet's water lilies in response to "Show me some famous paintings by Monet." Such an interaction model is contrasted with models for information retrieval (IR), where users are presented with a list of potentially relevant documents that they must then sort through manually.

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, unrestricted full-text understanding is beyond the state of the art in natural language processing, and furthermore, not all information is text; sounds, images, video, and other multimedia can all be valuable sources of knowledge. "Understanding" all these various media would require spectacular breakthroughs in other areas of artificial intelligence, such as object recognition, scene analysis, speech transcription, etc. In short, we are still years away from machines capable of distilling "meaning" from various types of multimedia documents.

Faced with the limitations of current technology and the insatiable thirst of users for more knowledge, what can we do? Rather than waiting for systems to be developed that can "understand" all available knowledge in various formats, 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? Our solution is natural language annotations [11], [12], [8], which are machine-parseable 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.

We have implemented this technology in START[1] [7], [8], the first natural language question answering system available on the World Wide Web.

To illustrate how our system works, consider the HTML fragment about Olympus Mons presented in Figure 1. It may be annotated with the following English sentences and phrases:

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

KIMAS 2003, October 1-3, 2003, Boston, MA, USA.

Mars' highest point
Largest volcano in the solar system
Olympus Mons is 25km tall.

START parses these annotations and stores the parsed structures (embedded ternary expressions [7]) with pointers back to the original information segment. To answer a question, the user query is compared against the annotations stored in the knowledge base. Because this match occurs at the level of syntactic structures, linguistically sophisticated machinery such as synonymy/hyponymy, ontologies, and structural transformation rules are all brought to bear on the matching process. Linguistic techniques allow the system to achieve capabilities beyond simple keyword matching, for example, handling complex syntactic alternations involving verb arguments. If a match is found between ternary expressions derived from annotations and those derived from the query, the segment corresponding to the annotations is returned to the user as the answer. For example, the annotations above allow START 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?

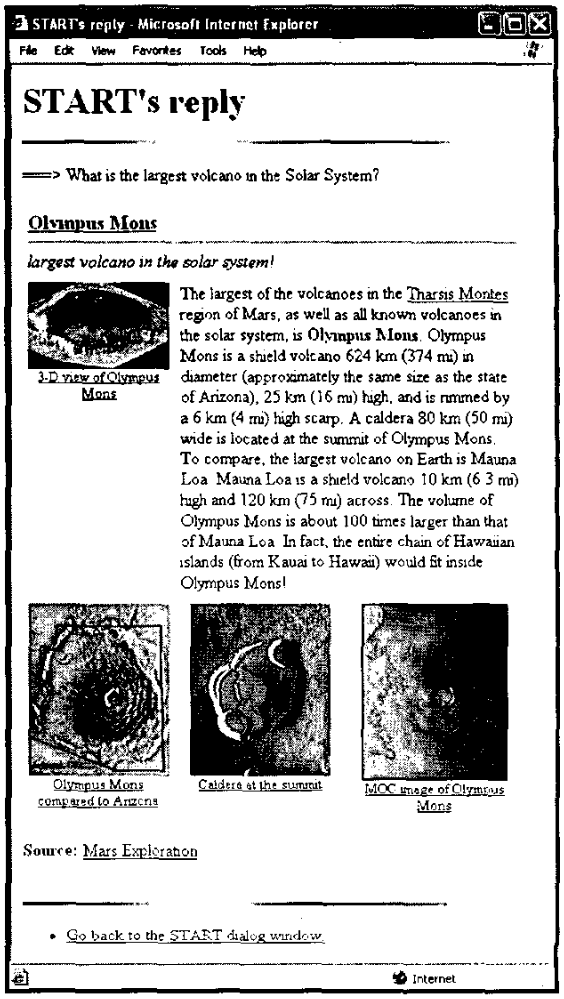
An important feature of the annotation concept is that any information segment can be annotated: not only text, but also images, multimedia, and even procedures! For example, multimedia items such as recordings of "hello" in various languages could be treated in the same manner (Figure 2). Pictures of famous people or flags of countries in the world could be annotated with appropriate phrases,and retrieved in response to user queries. A procedure for calculating distances between two locations or a procedure for calculating the current time in any world city could also be annotated for question answering.

Since it came on-line in December, 1993, START has engaged in exchanges with hundreds of thousands of users all over the world, supplying them with useful knowledge.

3. Annotations and Structured Knowledge

The ability to respond to natural language questions with textual and multimedia content crucially depends on natural language annotations. Because of this, the knowledge coverage of the START system is dependent on the amount of annotated material. To increase the effectiveness of our technology, we have adapted natural language annotations to work with structured and semistructured data.

If someone is asked a question like "When did Rutherford Hayes become president of the U.S.?", he or she might locate a resource with the answer—say, a book on famous people, or a Web site about presidents—find the section for Rutherford B. Hayes, and look up the date of his inauguration. Millions of questions can be answered by following this same recipe: extract an *object* (Rutherford Hayes) and a *property* (presidential term) from the question, find a data source (e.g., the POTUS Web site, `http://www.ipl.org/ref/POTUS`) for that type of object, look up the object's Web page, and extract the *value* for the answer (see Figure 3). By generalizing such plans and integrating them into a question answering system, we can achieve information access with broad coverage.



START's reply

==> What is the largest volcano in the Solar System?

Olympus Mons

largest volcano in the solar system!

3-D view of Olympus Mons

The largest of the volcanoes in the Tharsis Montes region of Mars, as well as all known volcanoes in the solar system, is **Olympus Mons**. Olympus Mons is a shield volcano 624 km (374 mi) in diameter (approximately the same size as the state of Arizona), 25 km (16 mi) high, and is rimmed by a 6 km (4 mi) high scarp. A caldera 80 km (50 mi) wide is located at the summit of Olympus Mons. To compare, the largest volcano on Earth is Mauna Loa Mauna Loa is a shield volcano 10 km (6 3 mi) high and 120 km (75 mi) across. The volume of Olympus Mons is about 100 times larger than that of Mauna Loa In fact, the entire chain of Hawaiian islands (from Kauai to Hawaii) would fit inside Olympus Mons!

Olympus Mons compared to Arizona

Caldera at the summit

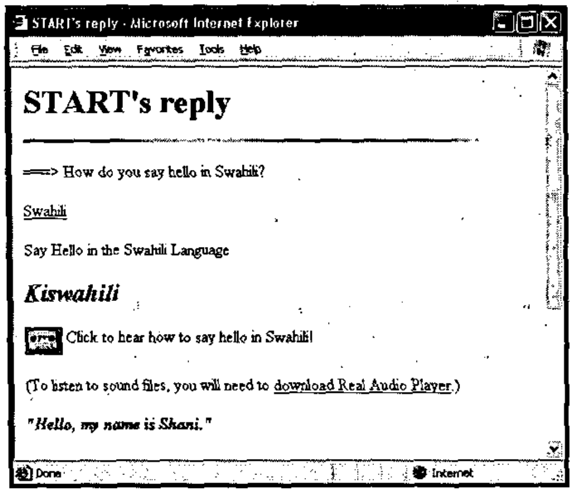
MOC image of Olympus Mons

Source: Mars Exploration

- Go back to the START dialog window.

Fig. 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.

The three main difficulties in getting a computer to answer such questions are understanding the question, identifying where to find the information, and fetching the information itself. START's parser is responsible for understanding user questions and translating them into structured queries. To help START address the other issues, we have developed a system called Omnibase[9], a "virtual" database that provides a uniform abstraction layer over multiple Web knowledge sources. Omnibase is capable of executing the structured queries generated by START. The following two sections will describe Omnibase in more detail.


START's reply - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Done
Internet

START's reply

==> How do you say hello in Swahili?

Swahili

Say Hello in the Swahili Language

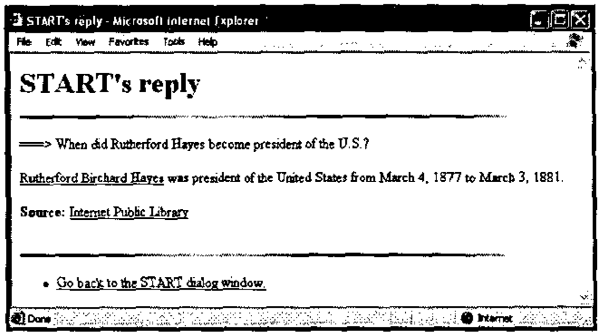
Kiswahili

Click to hear how to say hello in Swahili!

(To listen to sound files, you will need to download Real Audio Player.)

"Hello, my name is Shani."

Fig. 2. A multimedia response with sound.


START's reply - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Done
Internet

START's reply

==> When did Rutherford Hayes become president of the U.S.?

Rutherford Birchard Hayes was president of the United States from March 4, 1877 to March 3, 1881.

Source: Internet Public Library

- Go back to the START dialog window.

Fig. 3. The START system answering a question with data from Omnibase, presented within a generated sentence.

4. The Web as a Database

Although the Web is predominantly comprised of unstructured 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.

Omnibase serves as a structured query interface to heterogeneous data on the World Wide Web. It is of course impossible to impose any uniform schema on the entire Web. Instead, Omnibase adopts a stylized relational model which we call the "object–property–value" data 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". Figure 4 shows, however, that there are many alternative ways to ask for the value of a property of an object. Often, properties can be encoded in the arguments of verbs, e.g., the subject of the verb *invent* serves as the *inventor* property of a particular object. Adjectives can also be interpreted as properties, e.g., "how big" is understood as requesting the *area* property of a particular country.

Clearly, many other possible types of queries do not fall into the object–property–value model, such as questions about the relation between two objects (e.g., "How can I get from Boston to New York?").[2] 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 37% of TREC-9 and 47% of TREC-2001 questions from the QA track.

In addition, our experiments reveal a type of "Zipf's Law" of question distribution—a small fraction of question types account for a significant portion of all user information requests. Many questions ask for the same kind information, differing only in the specific object questioned, e.g., "Who directed Gone with the Wind?", "Who directed Star Wars?", "Who directed Good Will Hunting?", etc.; we can naturally group such questions together into a single question type, i.e., "Who directed *x*?" where *x* can be any movie. Such questions can be easily captured by our object–property–value data model. We find that structuring Web resources using this framework allows us to achieve relatively broad coverage with a modest number of different resources.

5. From Language to Knowledge

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.

Suppose the user asks "Who directed gone with the wind?" A natural language system cannot analyze this question without first knowing that "Gone with the Wind" can be treated as a single lexical item—otherwise, the question would make no more sense than, say, "Who hopped flown down the street?" Omnibase identifies the names of objects and the data sources they are associated with; for example, "Good Will Hunting" comes from a movie data source, "United States" comes from a country data source, etc. Not only does this help START understand the user question (which can now be read as "Who directed *X*"), but it also lets START know what data source contains the information, i.e., look in a movie database.

Since annotations can describe arbitrary fragments of knowledge, there is no reason why they can't be employed to describe Omnibase queries. In fact, annotations can be parameterized, i.e., they can contain symbols representative

[2] although START is capable of handling such questions in a more ad-hoc fashion

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]

Fig. 4. Some sample questions that can be handled by an object–property–value model of Web data.

of an entire class of objects. For example, the annotation "a person wrote the screenplay for `imdb-movie`" can be attached to an Omnibase query that retrieves the writers for various movies from the Internet Movie Database (IMDb). Note that because Omnibase has knowledge of movies, it can tell START which lexical items are actually movies. 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 in the Omnibase query.

Thus, with help from Omnibase, START translates user queries into a structured request (in the object–property–value model):

```
(get "imdb-movie"
     "Gone with the Wind (1939)"
     "DIRECTOR")
```

In this case, our natural language system needed to figure out that the user is asking about the DIRECTOR property of the object `"Gone with the Wind (1939)"`, and that this information can be found in the data source `imdb-movie`, corresponding to the Internet Movie Database.

Omnibase looks up the data source and property to find an associated script and applies the script to the object in order to retrieve the property value for the object.[3] The execution of the `imdb-movie` DIRECTOR script involves looking up a unique identifier for the movie (stored locally), fetching the correct page from the IMDb Web site (via a CGI interface), and matching a textual landmark on the page (literal text and HTML tags) to find the director of the movie. As a result, the list of movie directors is returned:

```
(get "imdb-movie"
     "Gone with the Wind (1939)"
     "DIRECTOR") =>
("George Cukor" "Victor Fleming" "Sam Wood")
```

Start then assembles the answer and presents it to the user either as a fragment of HTML or couched in natural language.

Currently, our system answers 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. 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 2000, to December 2002, about a million questions were posed to START and Omnibase. Of those, 67% were answered successfully by our system (59% of the questions answered were handled by Omnibase).

6. Large-Scale Syntactic Indexing

Although full syntactic and semantic analysis of open-domain natural language text is beyond current technology, we believe that it is possible to augment START's manual-annotation-based approach with automatically constructed annotations by extracting a limited subset of relations from unstructured text and using those relations to answer questions. In short, we advocate information retrieval on the level of key relations, in addition to keywords. This approach is promising because it attempts to address the well-known shortcomings of standard "bag-of-words" information retrieval techniques without requiring manual intervention.

The fragment pairs below illustrate the elusive nature of "meaning"; although fragments in each pair are nearly indistinguishable in terms of lexical content, their meanings are vastly different. Naturally, because one text fragment may be an appropriate answer to a question while the other fragment may not be, a question answering system seeking to achieve high precision must differentiate the semantic content of the pairs:

(1a) The bird ate the snake.
(1b) The snake ate the bird.
(2a) the largest planet's volcanoes
(2b) the planet's largest volcanoes
(3a) the house by the river
(3b) the river by the house
(4a) The Germans defeated the French.
(4b) The Germans were defeated by the French.

Ideally, question answering should be based on the semantics of questions and documents, but unfortunately, full semantic analysis is presently feasible only in highly restricted domains. Instead, we believe that a more pragmatic solution is to capture the relations relevant for question answering by automatically distilling natural language text into ternary expressions, like those used in START [7]. Such representations can easily express many types of relations, e.g., subject-verb-object relations, possession relations, etc. Using ternary expressions, the semantic differences between the text fragments presented above can be distinguished at the syntactic level:

[3] Such scripts are sometimes called wrappers [3].

(1a) [bird eat snake]
(1b) [snake eat bird]
(2a) [largest `adjmod` planet]
[planet `poss` volcano]
(2b) [largest `adjmod` volcano]
[planet `poss` volcano]
(3a) [house by river]
(3b) [river by house]
(4a) [Germans defeat French]
(4b) [French defeat Germans]

To test this idea, we have implemented Sapere, a prototype question answering system based on matching syntactic relations derived from the question with those derived from the corpus [15], [10]. We have evaluated Sapere against existing IR-based question answering systems using a restricted query set on an electronic version of the WorldBook Encyclopedia. To support our evaluation, we identified two linguistic phenomena, called *semantic symmetry* and *ambiguous modification*, that would benefit greatly from syntactic analysis [10].

Examples representing typical results from current question answering systems help illustrate the phenomena:

> **(Q1) What do frogs eat?**
>
> (A1) Adult *frogs eat* mainly insects and other small animals, including earthworms, minnows, and spiders.
>
> (A2) Alligators *eat* many kinds of small animals that live in or near the water, including fish, snakes, *frogs*, turtles, small mammals, and birds.
>
> (A3) Some bats catch fish with their claws, and a few species *eat* lizards, rodents, small birds, tree *frogs*, and other bats.
>
> **(Q2) What is the largest volcano in the Solar System?**
>
> (B1) Mars boasts many extreme geographic features; for example, Olympus Mons, the *largest volcano in the solar system.*
>
> (B2) The Galileo probe's mission to Jupiter, the *largest* planet *in the Solar system*, included amazing photographs of the *volcanoes* on Io, one of its four most famous moons.
>
> (B3) Even the *largest volcanoes* found on Earth are puny in comparison to others found around our own cosmic backyard, *the Solar System.*
>
> (B4) Olympus Mons, which spans an area the size of Arizona, is the *largest volcano in the Solar System.*

The first example (Q1) demonstrates the problem of semantic symmetry: although the questions "What do frogs eat?" and "What eats frogs?" are similar at the word level, they have very different meanings and should be answered differently. The second example (Q2) demonstrates the problem of ambiguous modification: adjectives like *largest* and prepositional phrases such as *in the Solar System* can modify a variety of different head nouns. Potential answers may contain the correct entities, but they may not be in the correct syntactic relations with each other, e.g., *the largest planet* instead of *the largest volcano.* Both these phenomena could benefit from a more detailed linguistic treatment to pinpoint more precise answers.

Semantic symmetry occurs when the selectional restrictions of different arguments of the same head overlap. For example, the selectional restriction for the subject of *eat* is *animate* and the selectional restriction for its object is *edible*; thus, semantic symmetry occurs whenever the subject and object of the verb *eat* are both animate and edible. In these cases, lexical content is insufficient to determine the meaning of the sentence—syntactic analysis is required to discover head-arguments relations.

Ambiguous modification occurs when an argument's selectional restrictions are so *un*restrictive that the argument can belong to more than one head in a particular context. Since nearly anything can be *large* or *good*, syntactic analysis is necessary to pin down which head this argument actually belongs to.

We have discovered that in answering questions that involved these phenomena, our relation-based approached demonstrated a significant increase in precision over standard keyword-based techniques. As an example, our baseline keyword-based system returned 32 results to the question "What eats frogs?" Of those, only one sentence actually answered the question (apparently, our poor frog has more predators than prey). Compare this to the results produced by Sapere:

> **(Q3) What do frogs eat?**
>
> (C1) Adult frogs eat mainly insects and other small animals, including earthworms, minnows, and spiders.

By examining subject-verb-object relations, Sapere can filter out irrelevant results and return only the correct responses.

7. Related Work

The use of natural language interfaces to access databases can be traced back to the sixties and seventies [4], [17], [6]; for a survey see [1]. Early research concentrated on adding natural language querying capabilities to existing relational databases. For the most part, the data was homogeneous and textual.

The idea of applying database techniques to the World Wide Web is not new. Many existing systems, e.g., ARANEUS [2], ARIADNE [14], Information Manifold [13], LORE [16], TSIMMIS [5], just to name a few, have attempted to unify heterogeneous 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 heterogeneous knowledge sources, the expressiveness of the query language, and implementation issues arising from the unreliable nature of the Web,[4] little work has been done on natural language querying capabilities.

What makes START and Omnibase unique among these systems is natural language question answering abilities and its use of the object–property–value data model. By allowing ordinary users to ask questions in English, we provide intuitive information access to a wealth of information. Furthermore, since our data model corresponds naturally to both user

[4] For a survey of database techniques for the Web, see [3].

questions and online content, the data integration task becomes more intuitive.

8. Conclusion

START and Omnibase are complementary components of a question answering system that addresses users' information access needs. START understands natural language questions and retrieves multimedia answers via annotations. Omnibase helps START translate natural language questions to structured queries. It serves as an abstraction layer which lets START treat heterogeneous Web sources as a uniform "virtual database". By providing a uniform natural language interface to heterogeneous knowledge on the World Wide Web, we can supply users with "just the right information." Finally, we can augment START's knowledge base with syntactic relations that are automatically derived from large amounts of natural language text. This technology expands the coverage of our knowledge bases without sacrificing precision.

Our natural language annotation technology, object–property–value data model, and relation extraction technology form a three-pronged approach that allows large-scale knowledge bases to be rapidly constructed from a variety of sources. By organizing and providing access to the World Wide Web and other resources, we hope to meet the future demands of information technology.

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

[2] Paolo Atzeni, Giansalvatore Mecca, and Paolo Merialdo. 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*, 1997.

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

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

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

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

[7] 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.

[8] 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.

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

[10] Boris Katz and Jimmy Lin. Selectively using relations to improve precision in question answering. In *Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering*, 2003.

[11] 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.

[12] 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.

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

[14] Craig Knoblock, Steven Minton, Jose Luis Ambite, Naveen Ashish, Ion Muslea, Andrew Philpot, and Sheila Tejada. The Ariadne approach to Web-based information integration. *International Journal on Cooperative Information Systems (IJCIS) Special Issue on Intelligent Information Agents: Theory and Applications*, 10(1/2):145–169, 2001.

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

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

[17] William A. Woods, Ronald M. Kaplan, and Bonnie L. Nash-Webber. The lunar sciences natural lanaugage information system: Final report. Technical Report 2378, BBN, 1972.