Information Access Using Natural Language

Boris Katz, Sue Felshin, Luciano Castagnola, Aaron Fernandes, Ali Ibrahim, Jimmy Lin,
Jerome McFarland, Alp Simsek & Baris Temelkuran

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: With recent advances in computer and Internet technology, people have access to more information than ever before. As the amount of information grows, so does the problem of finding what one is looking for. We believe that the most natural form of communication and information access for humans is natural language. We propose to address the growing information access problem with a uniform natural language interface.

Motivation: In December 1993, START (SynTactic Analysis using Reversible Transformations) became the first natural language system available for question answering on the World Wide Web. Since then START has been involved in dialogs with users all over the world, answering millions of questions. As we added more and more information to START's knowledge base, we discovered the advantages of "virtual collaboration." We realized that the existence of the Web, with its huge resources, allows us to put to use the fruits of labor of a large group of people without explicitly collaborating with them. Whenever we find an interesting knowledge source we can analyze its structure and interface it with START. Then, in response to a question, START can dispatch the user to a weather forecast page, a map collection, a World Factbook database, a personal homepage, etc., all accessed through our uniform natural language interface. It is this "virtual collaboration," which became possible only a few years ago, that inspired our project.

Previous Work: The project is based on the START natural language system developed by Boris Katz [4, 5]. The START system analyzes sentences as embedded *ternary expressions* which are indexed in a knowledge base. START has been used by researchers at MIT and other universities and research laboratories to construct and query knowledge bases using English [2]. To date, the virtual collaboration technique has enabled START to access a broad range of information in a number of topic areas, including corporations and their stock prices, weather reports, U.S. colleges and universities, U.S. presidents, movies, and more.

Approach: We use several approaches to integrate new knowledge sources into the START system. At the most direct level, if the source consists of sentences that can be processed by START (or if it can be converted into such a format), START reads the sentences and adds the information to its knowledge base. If the source consists of large amounts of unrestricted text, or is not amenable to language processing (such as pictures, maps, most Web pages, etc.), then we use an annotation mechanism [2] to label the pieces of information with phrases and sentences that START can understand. Finally, if access to the knowledge source requires some processing (as is the case in relational databases, searchable Web indexes and the like), START transforms a user's question into a query appropriate for the knowledge source. Our Omnibase program [3] makes the integration phase more seamless and uniform across different knowledge sources.

Some constructions found in text, such as numbers, dates, addresses and proper names, are better handled outside the conventional grammar/lexicon formalism, in a preprocessing stage which uses special purpose pattern recognizers based on real data (see [1]), or pre-calculated tables of known names (see [3]).

Although START is fairly robust in processing user queries, sometimes it cannot fully understand the sentence due to lexical or syntactic difficulties. Instead of faltering, START sometimes resort to less precise but more robust techniques of information retrieval, such as [6], to discover whether the relevant information is in START's knowledge base. Such techniques must be integrated with great care to ensure that they do not reduce the system's precision to an unacceptable level.

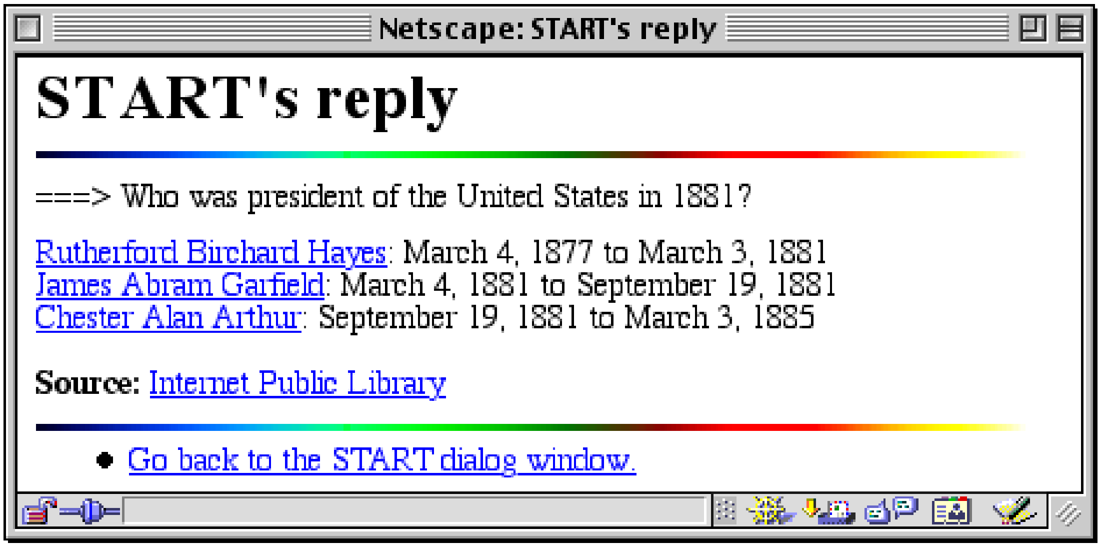
Difficulty: There are two main difficulties in integrating diverse knowledge sources into a single natural language interface. The first difficulty, common to any language processing system, comes from the complexity of the language itself. Despite decades of research, full processing of unrestricted text is still beyond reach. The second difficulty is one of integration. When a single system is able to access many knowledge sources simultaneously, the interference between sources becomes a serious problem.

Impact: A natural language system is the most intuitive interface for humans seeking information. It can produce high precision responses which decrease search time and increase productivity. Such a system requires less training, is accessible to a wider audience, and can be deployed in a shorter period of time; it can serve as a rigorous testbed for research in language understanding and multi-agent collaborative problem solving.

Future work: We will increase START's robustness, expand its syntactic and lexical coverage, and create a multi-agent knowledge access infrastructure.

Some knowledge sources are just too large to annotate manually. To solve this problem, we are working on partial parsing and more linguistically informed information retrieval techniques which will generate annotations automatically using those sentence fragments which were successfully analyzed by the system.

Once START correctly identifies the objects mentioned in the query, it needs to know what type of questions each knowledge source can answer. With Omnibase [3], a uniform interface for integrating multiple knowledge sources, START can access computer programs, inference engines, qualitative physics simulators, visual reasoning systems, etc. Solving the integration problem is central to the building of scalable AI systems of the future.



START's reply

===> Who was president of the United States in 1881?

Rutherford Birchard Hayes: March 4, 1877 to March 3, 1881
James Abram Garfield: March 4, 1881 to September 19, 1881
Chester Alan Arthur: September 19, 1881 to March 3, 1885

Source: Internet Public Library

- Go back to the START dialog window.

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

References:

[1] B. Katz et al. Blitz: A Preprocessor for Heuristically Detecting Context-Independent Linguistic Structures. In *MIT Artificial Intelligence Laboratory Research Abstracts (this volume)*, September 2001.

[2] B. Katz et al. From Sentence Processing to Multimedia Information Access. In *MIT Artificial Intelligence Laboratory Research Abstracts (this volume)*, September 2001.

[3] B. Katz et al. Omnibase: A Universal Data Source Interface. In *MIT Artificial Intelligence Laboratory Research Abstracts (this volume)*, September 2001.

[4] B. Katz. A Three-step Procedure for Language Generation. A.I. Memo 599, MIT AI Laboratory, 1980.

[5] B. Katz. Using English for Indexing and Retrieving. In P. H. Winston and S. A. Shellard, editors, *Artificial Intelligence at MIT: Expanding Frontiers*, volume 1, Cambridge, MA, 1990. MIT Press.

[6] B. Katz, J. Lin, and S. Felshin. Improving the Precision of Information Retrieval Systems Using Syntactic Relations. In *MIT Artificial Intelligence Laboratory Research Abstracts (this volume)*, September 2001.