From Sentence Processing to Multimedia Information Access

Boris Katz

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: It has become clear that a robust *full-text* natural language question-answering system cannot realistically be expected any time soon. Numerous problems such as intersentential reference and paraphrasing, summarization, common sense implication, and many more, will take a long time to solve to everybody's satisfaction. At the same time, however, it turns out that given a sophisticated grammar, a large lexicon enhanced by advances in Lexical Semantics, and an inference engine, it is possible to build a natural language system with satisfactory *sentence-level* performance.

Motivation: We need a mechanism that will let us bridge the gap between our ability to analyze natural language sentences and our appetite for processing huge amounts of natural language text and multimedia.

Previous Work: Our work builds on the START natural language system [1], which has been used by researchers at MIT and other universities and research laboratories for constructing and querying knowledge bases using English.[1] The START system analyzes English text and produces a *knowledge base* which incorporates, in the form of embedded *ternary expressions*, the information found in the text. One can think of the resulting entry in the knowledge base as a "digested summary" of the syntactic structure of an English sentence. A user can retrieve the information stored in the knowledge base by querying it in English. The system will then produce an English response.

A representation mimicking the hierarchical organization of natural language syntax has one undesirable consequence: sentences differing in their surface syntax but close in meaning are not considered similar by the system. START solves the problem by deploying *S-rules* (in forward and backward modes) which make explicit the relationship between alternate realizations of the arguments of verbs.

Approach: The START system attempts to bridge the gap between current sentence-level text analysis capabilities and the full complexity of unrestricted natural language by employing *natural language annotations* [2]. Annotations are computer-analyzable collections of natural language sentences and phrases that describe the contents of various information segments. START analyzes these annotations in the same fashion as any other sentences, but in addition to creating the required representational structures, the system also produces special pointers from these representational structures to the information segments summarized by the annotations.

Suppose, for example, that a user wants to retrieve this text fragment related to the discovery of Neptune:

> Neptune was discovered using mathematics. Before 1845, Uranus was widely believed to be the most distant planet. However, astronomers observed that Uranus was not always in the position predicted for it. The astronomers concluded that the gravitational attraction of a more distant planet was disturbing the orbit of Uranus. In 1845, John Adams, an English astronomer, calculated the location of this more distant planet. Urbain Leverrier, a French mathematician, independently did similar calculations. In 1846, John G. Galle and Heinrich d'Arrest of the Urania Observatory in Berlin, looked for the planet where Leverrier and Adams predicted it would be located. They saw the planet, which was later named Neptune, on September 23, 1846.

Let us assume that sentence (1) below serves as one of the annotations[2] to this text fragment:

(1) John Adams discovered Neptune using mathematics.

[1] For other approaches to the design of natural language querying systems, see, for example, [3].

[2] In the current version of the START system, most annotations are entered manually, although we are experimenting with several approaches that will make this process more automatic.

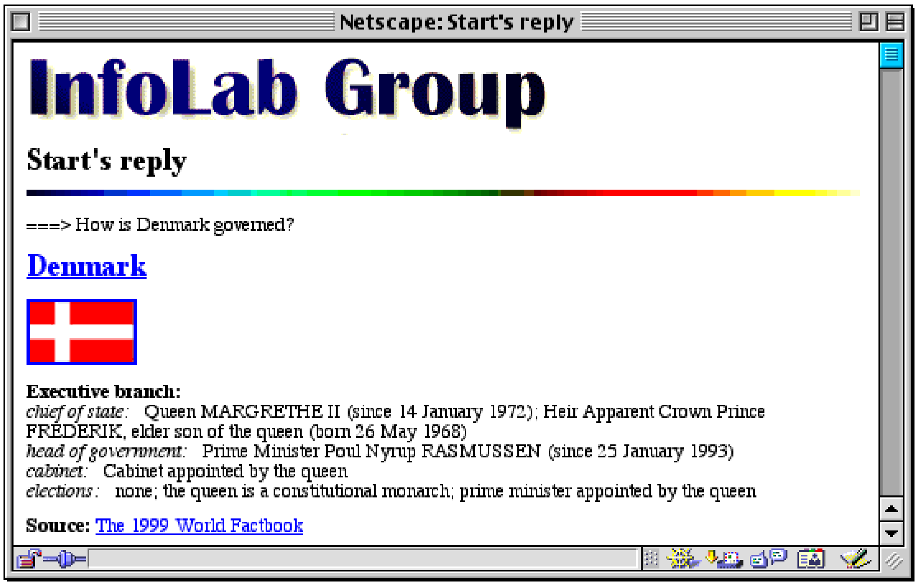
This means that START analyzed sentence (1) and incorporated it into the knowledge base along with a pointer to the text fragment. Now suppose the user asks one of the following questions:

(2) Who discovered Neptune?
Did Adams discover Neptune?
How was Neptune discovered?
Was Neptune discovered using mathematics?
Tell me about Neptune's discovery.

START begins the process of answering any such question by creating a ternary expression to be matched against the knowledge base. It is important to emphasize that the full power of sentence-level natural language processing is brought to bear on the matching process. START's matcher works both on the *word* level (using, if appropriate, additional lexical information about synonyms, hyponyms, IS-A trees, *etc.*) and on the *structure* level (utilizing necessary S-rules, information on verb-class membership, nominalization *etc.*), although in the case of very simple questions such as (2) most of this machinery is not utilized.

Since the representational structure returned by the matcher contains a special pointer to the annotated text fragment, START's usual sentence-level question-answering strategy is modified. Instead of passing the representational structure to the language generation system and asking it to produce an English sentence such as (1), START simply follows the pointer and presents the text fragment to the user. The annotation technique has been used to create START knowledge bases concerning geographical information about cities and countries of the world, research at the MIT AI Laboratory, famous people, colleges, movies, and more. Most recently, we have developed a technique of nesting annotations within other annotated material, allowing us to perform comparisons and other complex retrieval operations more easily.

Impact: The natural language annotation technique easily generalizes to the indexing and retrieval of all types of information, whether text-based or not. Using START, one can immediately access text, images, sound, video, Web pages, and more without relying on the as-yet-unreached goal of full language processing.



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. Using English for Indexing and Retrieving. in P. H. Winston and S. A. Shellard (eds.), *Artificial Intelligence at MIT: Expanding Frontiers,* vol. 1, MIT Press, 1990.

[2] B. Katz. "Annotating the World Wide Web using Natural Language." Proceedings of RIAO'97, Montreal, Canada, 1997.

[3] J.F. Allen and L.K. Schubert "The TRAINS project." Technical Report 382, Department of Computer Science, University of Rochester, Rochester, N.Y. 1991.