Annotating the World Wide Web

Boris Katz, Jimmy Lin & Sue Felshin

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Although vast amounts of information are available electronically today, no effective mechanism exists to provide humans with convenient access to that information.

Motivation: Keyword search engines are popular because they provide results—often, too many results! If we could understand, at least partially, the *meaning* of documents, rather than just recognizing the *words*, we could answer queries with much higher precision. Natural language is the most convenient and most intuitive method of information access, and people should be able to retrieve information using a system capable of understanding and answering natural language questions.

Previous Work: In [2], we proposed making use of natural language annotations to annotate the World Wide Web.

Approach: Technology is not up to the task of analyzing the semantic content of unrestricted data—complex text, tables, images, sound files, etc. However, our experiments with the START system [3] show how this problem could be solved for a relatively small knowledge base using our annotation technology. Annotations are short, simple sentences and phrases which computers *can* analyze. We associate them with data which is opaque to analysis. Our question answering system can then analyze questions, match them against already analyzed and stored annotations, and retrieve opaque data associated with matching annotations. For example, we can associate an individual annotation, such as the sentence "John Adams discovered Neptune using mathematics," with a detailed paragraph concerning the discovery of Neptune; see [1]. Figure 1 shows the process of creating annotations and retrieving information.

Much online data is stored in databases with regular format. Rather than writing annotations for each database entry, we can write "parameterized annotations" which access database entries of the same type. For example, the annotation "*any-IMDb-person* directs *any-IMDb-movie*" can be used to answer queries about the director of any movie in the Internet Movie Database.

Even with the use of parameterized annotations, there is still a great deal of manual labor involved in writing annotations in order to build a system with a useful amount of knowledge. The solution to this problem is to get everyone involved. For a limited dataset such as corporate data (e.g., "organizational memory"), members of the organization can be asked to contribute annotations. But for a large, heterogeneous, and uncontrolled dataset such as the World Wide Web, we must turn to volunteers. Large numbers of individuals have shown a willingness to volunteer their efforts, even without the promise of personal gain, as shown by projects such as the Open Mind Initiative [4, 5], which is a recent effort to organize ordinary users on the World Wide Web (*netizens*) to assist in developing intelligent software, Open Mind Commonsense[5], which is an attempt at constructing a large common sense database by collecting assertions from users all over the Web, and *dmoz*, the Open Directory Project[6], whose goal is to produce the most comprehensive directory of the Web by relying on volunteer editors.

Difficulties remain in controlling the quality of annotations. These problems could be addressed through the use of editors, as in the Open Source movement, which promotes software development by nurturing a community of individual contributors working on freely distributed source code. Under this development model, software reliability and quality is ensured through independent peer review by a large number of programmers.

If the system is applied to a large dataset such as the World Wide Web, there are questions regarding system coverage. Before a critical mass of knowledge is attained, users' expectations will be carefully managed so that they

[5] `http://openmind.media.mit.edu`
[6] `http://www.dmoz.org`

realize the system is highly experimental and has a very limited range of knowledge. It may be necessary to restrict the initial system to limited domains of knowledge, and only gradually expand the domains, to ensure that users can have a reasonable expectation of receiving an answer to their query.

Impact: Through sufficient volunteer effort in creating annotations, we could provide users with far more effective and more convenient search capabilities on the World Wide Web or within smaller datasets.

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

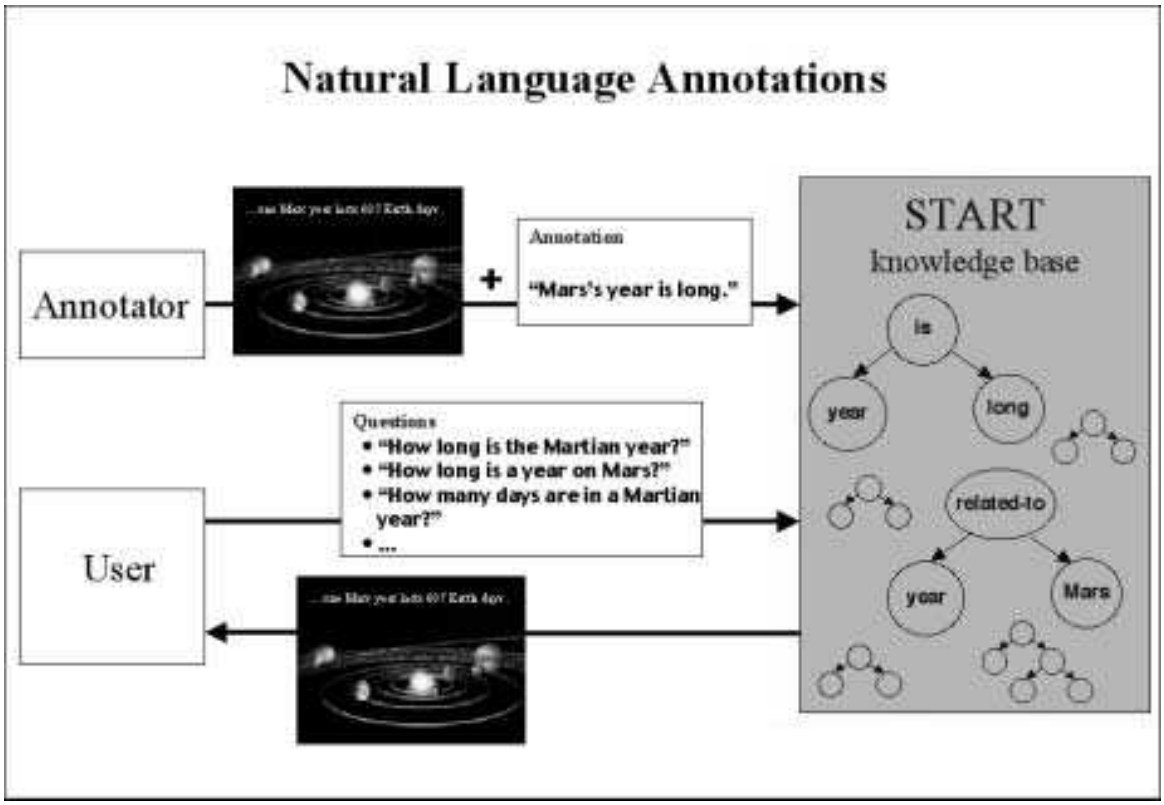

Figure 1: Natural language annotations

References:

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

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

[3] B. Katz, S. Felshin, L. Castagnola, A. Fernandes, A. Ibrahim, J. Lin, J. McFarland, A. Simsek, and B. Temelkuran. Information Access Using Natural Language. In *MIT Artificial Intelligence Laboratory Research Abstracts (this volume)*, September 2001.

[4] David G. Stork. Character and Document Research in the Open Mind Initiative. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, 1999.

[5] David G. Stork. Open data collection for training intelligent software in the Open Mind Initiative. In *Proceedings of the Engineering Intelligent Systems Symposium (EIS '2000)*, 2000.