

Linguistic Knowledge Can Improve Information Retrieval

William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, and Stephen Green

Introduction by William A. Woods

This article was published in the Proceedings of the Applied Natural Language Processing Conference (ANLP-2000) in Seattle, Washington, May 1-3, 2000. A preliminary version was published as a technical report (TR-99-83) in the Sun Microsystems Laboratories Technical Report Series. The article represents a milestone in an ongoing project aimed at discovering technology to help people find specific information online. We undertook this project with the understanding that if we could significantly improve people's ability to find information online, then this would have wide impact across a variety of computer applications, from finding information in online documentation, to finding needed information on a Web site; from following ideas in research material, to organizing case material for legal proceedings; from supporting telephone call centers, to publishing information with indexes that enable people to efficiently find what they need; and on and on ... This project anticipated the needs that many people have now experienced using search engines on the Web.

We began by looking closely at what goes wrong when people try to find information online and looking for ways to solve the problems that we identified. One problem that immediately became apparent was that people often asked questions using different terminology from that used by the author of the material they needed to find. Sometimes they used different morphological forms of a word, like "compute" versus "computation," and sometimes they used different semantically related words, like "go" versus "move." We've come to call this the "paraphrase problem." Another obvious problem was that people spent an inordinate amount of time looking through the documents returned by a request, trying to determine whether they actually contained the information that was needed (and unfortunately, most of them didn't).

This investigation led to the discovery of the techniques described in this article, in which the indexing system automatically constructs a conceptual taxonomy of all the words and phrases in the indexed material, organized by generality, using an extensive body of linguistic knowledge, including knowledge of semantic relationships among concepts and morphological structure and relationships between words. We call this "conceptual indexing." This taxonomy is then used by a specific passage retrieval algorithm to make connections between what you ask for and what you need to find. The taxonomy also supports very effective deep browsing. The specific passage retrieval system locates not just documents, but specific passages in those documents that are likely to contain the information you need, and it effectively ranks these passages using a scoring method that really does tend to rank the best passages first. The system saves people time spent searching and improves the quality of decisions, using linguistic knowledge to help them find what they really want.

This project is an activity of the Knowledge Technology Group in Sun Microsystems Laboratories, whose goal is to develop and exploit technology for dealing with knowledge -- acquiring, organizing, disseminating, retrieving, and browsing it. Current members of the team are: William A. Woods, Stephen Green, Paul Martin, Robert J. Kuhns, Ann Houston, and summer intern, Scott Sanner. Former members are acknowledged in the article. The project has developed a powerful technology that is now used internally at Sun in a number of applications, and some of it has been included in Sun products. Going forward, we continue to look for ways to exploit what has been developed so far and to discover new ways to help people deal with knowledge online.

You can find a brief overview of conceptual indexing technology at:
<http://www.sun.com/research/knowledge/>

You can use the technology to search the Sun Microsystems Laboratories Web site at:
<http://www.research.sun.com/>

REFERENCES:

"Conceptual Indexing: A Better Way to Organize Knowledge," W. A. Woods, Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April, 1997.

Online at: <http://www.sun.com/techrep/1997/abstract-61.html>

"Finding the Hidden Gems: Search Engines and Beyond," W. A. Woods, Knowledge Management in the Telecoms Industry, First Conferences, Ltd., London, ENGLAND, October 23-24, 1997.

"Knowledge Management Needs Effective Search Technology," W. A. Woods, Sun Journal, March, 1998. Online at: http://www.sun.com/sun-journal/V2N1/03_feat2a.html

"Natural Language Technology in Precision Content Retrieval," J. Ambroziak and W. A. Woods, proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA 98), August 18-21, 1998, Moncton, New Brunswick, CANADA.

Online at: <http://www.sunlabs.com/techrep/1998/abstract-69.html>

"Linguistic Knowledge can Improve Information Retrieval," William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, and Stephen Green, Proceedings of ANLP-2000, Seattle, WA, May 1-3, 2000, (preliminary version: Technical Report SMLI TR-99-83, Sun Microsystems Laboratories, Mountain View, CA, December, 1999.

Online at: <http://www.sun.com/research/techrep/1999/abstract-83.html>

"Aggressive Morphology for Robust Lexical Coverage," William A. Woods, Proceedings of ANLP-2000, Seattle, WA, May 1-3, 2000, (preliminary version: Technical Report SMLI TR-99-82, Sun Microsystems Laboratories, Mountain View, CA, December, 1999.

Online at: <http://www.sun.com/research/techrep/1999/abstract-82.html>

"Conceptual Indexing: Practical Large-Scale AI for Efficient Information Access," William A. Woods, Invited talk at AAAI 2000, Proceedings AAAI 2000, Austin, TX, August 2, 2000.

"Halfway to Question Answering," by W. A. Woods, Stephen Green, Paul Martin, and Ann Houston, Proceedings of the TREC-9 Conference, November, 13-16, 2000.