

Demonstration of Precision Content Retrieval

Stephen Green, Paul Martin, and William A. Woods

Sun Microsystems Laboratories

1 Network Drive

Burlington, MA 01803

William.Woods@east.sun.com

Abstract

We will demonstrate a linguistically augmented conceptual indexing and retrieval system that finds relevant passages in indexed text material in response to specific information requests. It uses syntactic, semantic, and morphological knowledge to automatically organize words and phrases from the indexed material into a taxonomy of concepts, ordered by generality, which is then used to make connections between terms in a request and terms that should be sought in relevant passages of text. The system uses a combination of linguistic content processing and intensional subsumption logic (Woods, 1991) to automatically construct a *conceptual index* (Woods, 1997) of all the words and phrases that occur in a body of text, organized by a relationship of generality (subsumption).

1 Description

Nearly everyone is familiar with the experience of searching the Web with a Web search engine and with using a search interface to search a particular web site once you get there. (You may even have noticed that the latter often doesn't work as well as the former.) After you have a list of hits, you may then spend a significant amount of time following links, waiting for pages to download, reading through the page to see if it has what you want, discovering that it doesn't, backing up and trying another link, deciding to try another way to phrase your request, etc., before finding what you want (or giving up and deciding that you can't find it).

A research project at Sun Microsystems Laboratories has developed a technology that completely changes this experience. Referred to as "Precision Content Retrieval", this technology combines linguistic knowledge and natural language processing techniques with a technique for "Conceptual Indexing" and a technique for "Specific Passage Retrieval" to take you directly to relevant passages of material that are likely to contain the answers to your questions, short circuiting most of that time spent clicking and downloading and reading and rephrasing.

To support this, the system automatically con-

structs a structured taxonomy of all of the concepts (i.e., meaningful words and phrases) that it finds in the text material when it indexes it, and it organizes this taxonomy by generality so that more general terms subsume more specific ones. This is the phase of "Conceptual Indexing". The resulting conceptual taxonomy can then be used for browsing and navigation and to help the system connect terms in a request to related terms that might be used instead in the material that you need to find. For example, using linguistic knowledge from its lexicon, the system can infer that "becomes black" is subsumed by "color change" in its conceptual taxonomy, so that a user's request for the latter will automatically find the former (as well as "reset bitmap colors", "color disruption", etc.).

When an information seeker asks a question, the system not only finds documents containing the requested concepts, but identifies the specific passages within those documents that contain the information requested, ranking them according to how exactly they match what was requested and displaying the actual passages in decreasing order of their likelihood of being useful. This is the phase of "Specific Passage Retrieval". The combination of these two techniques results in a major breakthrough in what previously seemed possible for online search effectiveness. In one experiment with this technology, a five-fold improvement in human search productivity was achieved, compared to using traditional document retrieval technology for the same task.

The experimental result cited above is described in an article "Linguistic Knowledge can Improve Information Retrieval," by William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, and Stephen Green (Woods et al., 2000).

You can get a brief overview of this technology at:

<http://www.sun.com/research/knowledge/>

This technology was the subject of a Keynote address at the AAAI-2000 and two conference papers in ANLP-2000 (Woods, 2000; Woods et al., 2000). It was first publicly demonstrated at this year's JavaOne(sm) conference.

2 Conclusion

We have implemented a system combining taxonomic subsumption techniques, linguistic knowledge, and natural language processing techniques with a penalty-based, relaxation-ranking, passage-retrieval algorithm to locate specific information in unrestricted text. This is a different approach from previous methods of passage retrieval and from previous attempts to use linguistic knowledge in information retrieval. Experiments have shown that this methodology can significantly improve information retrieval performance and human search productivity. Of particular interest is the way that morphological and syntactic analysis and semantic subsumption are integrated in the conceptual taxonomy and the way that this information interacts with actual requests and real data.

Acknowledgments

Many other people have been involved in creating the conceptual indexing and retrieval system described here. These include: Gary Adams, Jacek Ambroziak, Lawrence Bookman, Chris Colby, Jim Flowers, Ellen Hays, Robert Kuhns, Patrick Martin, Peter Norvig, Tony Passera, Philip Resnik, Scott Sanner, Robert Sproull, and Mark Torrance.

Sun, Sun Microsystems, and JavaOne are trademarks or registered trademarks of Sun Microsystems, Inc. in the U.S. and other countries. Sun, Sun Microsystems, et JavaOne sont des marques dposes ou enregistres de Sun Microsystems, Inc. aux Etats-Unis et dans d'autres pays.

References

- William A. Woods, Lawrence A. Bookman, Ann C. Houston, Robert J. Kuhns, Paul A. Martin, and Stephen Green. 2000. Linguistic knowledge can improve information retrieval. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, Seattle WA, April. ACL ANLP-2000. available at: <http://research.sun.com/research/features/tenyears/volcd/papers/woods.htm>.
- William A. Woods. 1991. Understanding subsumption and taxonomy: A framework for progress. In John Sowa, editor, *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, pages 45–94. Morgan Kaufmann, San Mateo, CA.
- William A. Woods. 1997. Conceptual indexing: A better way to organize knowledge. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April. www.sun.com/research/techrep/1997/abstract-61.html.
- William A. Woods. 2000. Aggressive morphology for robust lexical coverage. In *Proceedings of the*