# Linguistic Knowledge can Improve Information Retrieval

**William A. Woods** and **Lawrence A. Bookman*** and **Ann Houston** and **Robert J. Kuhns** and **Paul Martin** and **Stephen Green**
Sun Microsystems Laboratories
1 Network Drive
Burlington, MA 01803
{William.Woods,Ann.Houston,Robert.Kuhns,Paul.Martin,Stephen.Green}@east.sun.com

## Abstract

This paper describes the results of some experiments using a new approach to information access that combines techniques from natural language processing and knowledge representation with a penalty-based technique for relevance estimation and passage retrieval. Unlike many attempts to combine natural language processing with information retrieval, these results show substantial benefit from using linguistic knowledge.

## 1 Introduction

An online information seeker often fails to find what is wanted because the words used in the request are different from the words used in the relevant material. Moreover, the searcher usually spends a significant amount of time reading retrieved material in order to determine whether it contains the information sought. To address these problems, a system has been developed at Sun Microsystems Laboratories (Ambroziak and Woods, 1998) that uses techniques from natural language processing and knowledge representation, with a technique for dynamic passage selection and scoring, to significantly improve retrieval performance. This system is able to locate specific passages in the indexed material where the requested information appears to be, and to score those passages with a penalty-based score that is highly correlated with the likelihood that they contain relevant information. This ability, which we call "Precision Content Retrieval" is achieved by combining a system for Conceptual Indexing with an algorithm for Relaxation-Ranking Specific Passage Retrieval.

In this paper, we show how linguistic knowledge is used to improve search effectiveness in this system. This is of particular interest, since many previous attempts to use linguistic knowledge to improve information retrieval have met with little or mixed success (Fagan, 1989; Lewis and Sparck Jones, 1996; Sparck Jones, 1998; Varile and Zampolli, 1997; Voorhees, 1993; Mandala et al., 1999) (but see the latter for some successes as well).

* Lawrence Bookman is now at Torrent Systems, Inc.

## 2 Conceptual Indexing

The conceptual indexing and retrieval system used for these experiments automatically extracts words and phrases from unrestricted text and organizes them into a semantic network that integrates syntactic, semantic, and morphological relationships. The resulting conceptual taxonomy (Woods, 1997) is used by a specific passage-retrieval algorithm to deal with many paraphrase relationships and to find specific passages of text where the information sought is likely to occur. It uses a lexicon containing syntactic, semantic, and morphological information about words, word senses, and phrases to provide a base source of semantic and morphological relationships that are used to organize the taxonomy. In addition, it uses an extensive system of knowledge-based morphological rules and functions to analyze words that are not already in its lexicon, in order to construct new lexical entries for previously unknown words (Woods, 2000). In addition to rules for handling derived and inflected forms of known words, the system includes rules for lexical compounds and rules that are capable of making reasonable guesses for totally unknown words.

A pilot version of this indexing and retrieval system, implemented in Lisp, uses a collection of approximately 1200 knowledge-based morphological rules to extend a core lexicon of approximately 39,000 words to give coverage that exceeds that of an English lexicon of more than 80,000 base forms (or 150,000 base plus inflected forms). Later versions of the conceptual indexing and retrieval system, implemented in C++, use a lexicon of approximately 150,000 word forms that is automatically generated by the Lisp-based morphological analysis from its core lexicon and an input word list. The base lexicon is extended further by an extensive name dictionary and by further morphological analysis of unknown words at indexing time. This paper will describe some experiments using several versions of this system. In particular, it will focus on the role that the linguistic knowledge sources play in its operation.

The lexicon used by the conceptual indexing system contains syntactic information that can be used

for the analysis of phrases, as well as morphological and semantic information that is used to relate more specific concepts to more general concepts in the conceptual taxonomy. This information is integrated into the conceptual taxonomy by considering base forms of words to subsume their derived and inflected forms ("root subsumption") and more general terms to subsume more specific terms. The system uses these relationships as the basis for inferring subsumption relationships between more general phrases and more specific phrases according to the intensional subsumption logic of Woods (Woods, 1991).

The largest base lexicon used by this system currently contains semantic subsumption information for something in excess of 15,000 words. This information consists of basic "kind of" and "instance of" information such as the fact that *book* is a kind of *document* and *washing* is a kind of *cleaning*. The lexicon also records morphological roots and affixes for words that are derived or inflected forms of other words, and information about different word senses and their interrelationships. For example, the conceptual indexing system is able to categorize *becomes black* as a kind of *color change* because *becomes* is an inflected form of *become*, *become* is a kind of *change*, and *black* is a *color*. Similarly, *color disruption* is recognized as a kind of *color change*, because the system recognizes *disruption* as a derived form of *disrupt*, which is known in the lexicon to be a kind of *damage*, which is known to be a kind of *change*.

When using root subsumption as a technique for information retrieval, it is important to have a core lexicon that knows correct morphological analyses for words that the rules would otherwise analyze incorrectly. For example, the following are some examples of words that could be analyzed incorrectly if the correct interpretations were not specified in the lexicon:

**delegate** (de+leg+ate) take the legs from

**caress** (car + ess) female car

**cashier** (cashy + er) more wealthy

**daredevil** (dared + evil) serious risk

**lacerate** (lace + rate) speed of tatting

**pantry** (pant + ry) heavy breathing

**pigeon** (pig + eon) the age of peccaries

**ratify** (rat + ify) infest with rodents

**infantry** (infant + ry) childish behavior

Although they are not always as humorous as the above examples, there are over 3,000 words in the core lexicon of 39,000 English words that would receive false morphological analyses like the above examples, if the words were not already in the lexicon.

## 3 Relaxation Ranking and Specific Passage Retrieval

The system we are evaluating uses a technique called "relaxation ranking" to find specific passages where as many as possible of the different elements of a query occur near each other, preferably in the same form and word order and preferably closer together. Such passages are ranked by a penalty score that measures the degree of deviation from an exact match of the requested phrase, with smaller penalties being preferred. Differences in morphological form and formal subsumption of index terms by query terms introduce small penalties, while intervening words, unexplained permutations of word order, and crossing sentence boundaries introduce more significant penalties. Elements of a query that cannot be found nearby introduce substantial penalties that depend on the syntactic categories of the missing words.

When the conceptual indexing system is presented with a query, the relaxation-ranking retrieval algorithm searches through the conceptual taxonomy for appropriately related concepts and uses the positions of those concepts in the indexed material to find specific passages that are likely to address the information needs of the request. This search can find relationships from base forms of words to derived forms and from more general terms to more specific terms, by following paths in the conceptual taxonomy.

For example, the following is a passage retrieved by this system, when applied to the UNIX® operating system online documentation (the "man pages"):

> Query: print a message from the mail tool
>
> 6. -2.84 **print mail mail mailtool**
>
> Print sends copies of all the selected mail items to your default printer. If there are no selected items, mailtool sends copies of those items you are currently...

The indicated passage is ranked 6th in a returned list of found passages, indicated by the 6 in the above display. The number -2.84 is the penalty score assigned to the passage, and the subsequent words *print*, *mail*, *mail*, and *mailtool* indicate the words in the text that are matched to the corresponding content words in the input query. In this case, *print* is matched to *print*, *message* to *mail*, *mail* to *mail*, and *tool* to *mailtool*, respectively. This is followed by the content of the actual passage located. The information provided in these hit displays gives the information seeker a clear idea of why the passage was retrieved and enables the searcher to quickly skip down the hit list with little time spent looking at irrelevant passages. In this case, it was easy to

identify that the 6th ranked hit was the best one and contained the relevant information.

The retrieval of this passage involved use of a semantic subsumption relationship to match *message* to *mail*, because the lexical entry for *mail* recorded that it was a kind of *message*. It used a morphological root subsumption to match *tool* to *mailtool* because the morphological analyzer analyzed the unknown word *mailtool* as a compound of *mail* and *tool* and recorded that its root was *tool* and that it was a kind of *tool* modified by *mail*. Taking away the ability to morphologically analyze unknown words would have blocked the retrieval of this passage, as would eliminating the lexical subsumption entry that recorded *mail* as a kind of *message*.

Like other approaches to passage retrieval (Kaszkiel and Zobel, 1997; Salton et al., 1993; Callan, 1994), the relaxation-ranking retrieval algorithm identifies relevant passages rather than simply identifying whole documents. However, unlike approaches that involve segmenting the material into paragraphs or other small passages before indexing, this algorithm dynamically constructs relevant passages in response to requests. When responding to a request, it uses information in the index about positions of concepts in the text to identify relevant passages. In response to a single request, identified passages may range in size from a single word or phrase to several sentences or paragraphs, depending on how much context is required to capture the various elements of the request.

In a user interface to the specific passage retrieval system, retrieved passages are reported to the user in increasing order of penalty, together with the rank number, penalty score, information about which target terms match the corresponding query terms, and the content of the identified passage with some surrounding context as illustrated above. In one version of this technology, results are presented in a hypertext interface that allows the user to click on any of the presented items to see that passage in its entire context in the source document. In addition, the user can be presented with a display of portions of the conceptual taxonomy related to the terms in the request. This frequently reveals useful generalizations of the request that would find additional relevant information, and it also conveys an understanding of what concepts have been found in the material that will be matched by the query terms. For example, in one experiment, searching the online documentation for the Emacs text editor, the request *jump to end of file* resulted in feedback showing that *jump* was classified as a kind of *move* in the conceptual taxonomy. This led to a reformulated request, *move to end of file*, which successfully retrieved the passage *go to end of buffer*.

## 4 Experimental Evaluation

In order to evaluate the effectiveness of the above techniques, a set of 90 queries was collected from a naive user of the UNIX operating system, 84 of which could be answered from the online documentation known as the man pages. A set of "correct" answers for each of these 84 queries was manually determined by an independent UNIX operating system expert, and a snapshot of the man pages collection was captured and indexed for retrieval. In order to compare this methodology with classical document retrieval techniques, we assign a ranking score to each document equal to the ranking score of the best ranked passage that it contains.

In rating the performance of a given method, we compute average recall and precision values at 10 retrieved documents, and we also compute a "success rate" which is simply the percentage of queries for which an acceptable answer occurs in the top ten hits. The success rate is the principal factor on which we base our evaluations, since for this application, the user is not interested in subsequent answers once an acceptable answer has been found, and finding one answer for each of two requests is a substantially better result than finding two answers to one request and none for another.

These experiments were conducted using an experimental retrieval system that combined a Lisp-based language processing stage with a C++ implementation of a conceptual indexer. The linguistic knowledge sources used in these experiments included a core lexicon of approximately 18,000 words, a substantial set of morphological rules, and specialized morphological algorithms covering inflections, prefixes, suffixes, lexical compounding, and a variety of special forms, including numbers, ordinals, Roman numerals, dates, phone numbers, and acronyms. In addition, they made use of a lexical subsumption taxonomy of approximately 3000 lexical subsumption relations, and a small set of semantic entailment axioms (e.g., *display* entails *see*, but is not a kind of *see*). This system is described in (Woods, 1997). The database was a snapshot of the local man pages (frozen at the time of the experiment so that it wouldn't change during the experiment), consisting of approximately 1800 files of varying lengths and constituting a total of approximately 10 megabytes of text.

Table 1 shows the results of comparing three versions of this technology with a textbook implementation of the standard *tf·idf* algorithm (Salton, 1989) and with the SearchIt™search application developed at Sun Microsystems, Inc., which combines a

Table 1: A comparison of different retrieval techniques.

| System | Success Rate | Recall (10 docs) | Precision (10 docs) |
|---|---|---|---|
| *tf·idf* | 28.6% | 14.8% | 2.9% |
| SearchIt system | 44.0% | 28.5% | 7.4% |
| Recall II | 60.7% | 38.6% | 7.3% |
| w/o morph | 50.0% | not measured | not measured |
| w/o knowledge | 42.9% | not measured | not measured |

simple morphological query expansion with a state-of-the-art commercial search engine. In the table, Recall II refers to the full conceptual indexing and search system with all of its knowledge sources and rules. The line labeled "w/o morph" refers to this system with its dynamic morphological rules turned off, and the line labeled "w/o knowledge" refers to this system with all of its knowledge sources and rules turned off. The table presents the success rate and the measured recall and precision values for 10 retrieved documents. We measured recall and precision at the 10 document level because internal studies of searching behavior had shown that users tended to give up if an answer was not found in the first ten ranked hits. We measured success rate, rather than recall and precision, for our ablation studies, because standard recall and precision measures are not sensitive to the distinction between finding multiple answers to a single request versus finding at least one answer for more requests.

## 5 Discussion

Table 1 shows that for this task, the relaxation-ranking passage retrieval algorithm without its supplementary knowledge sources (Recall II w/o knowledge) is roughly comparable in performance (42.9% versus 44.0% success rate) to a state-of-the-art commercial search engine (SearchIt) at the pure document retrieval task (neglecting the added benefit of locating the specific passages). Adding the knowledge in the core lexicon (which includes morphological relationships, semantic subsumption axioms, and entailment relationships), but without morphological analysis of unknown words (Recall II w/o morph), significantly improves these results (from 42.9% to 50.0%). Further adding the morphological analysis capability that automatically analyzes unknown words (deriving additional morphological relationships and some semantic subsumption relationships) significantly improves that result (from 50.0% to 60.7%). In contrast, we found that adding the same semantic subsumption relationships to the commercial search engine, using its provided thesaurus capability degraded its results, and results were still degraded when we added only those facts that we knew would help find relevant documents. It turned out that the additional relevant documents found were more than offset by additional irrelevant documents that were also ranked more highly.

## 6 Anecdotal Evaluation of Specific Passage Retrieval Benefits

As mentioned above, comparing the relaxation-ranking algorithm with document retrieval systems measures only a part of the benefit of the specific passage retrieval methodology. Fully evaluating the quality and ranking of the retrieved passages involves a great many subtleties. However, two informal evaluations have been conducted that shed some light on the benefits.

The first of these was a pilot study of the technology at a telecommunications company. In that study, one user found that she could use a single query to the conceptual indexing system to find both of the items of information necessary to complete a task that formerly required searching two separate databases. The conclusion of that study was that the concept retrieval technology performs well enough to be useful to a person talking live with a customer. It was observed that the returned hits can be compared with one another easily and quickly by eye, and attention is taken directly to the relevant content of a large document. The automatic indexing was considered a plus compared with manual methods of content indexing. It was observed that an area of great potential may be in a form of knowledge management that involves organizing and providing intelligent access to small, unrelated "nuggets" of textual knowledge that are not amenable to conventional database archival or categorization.

A second experiment was conducted by the Human Resources Webmaster of a high-tech company, an experienced user of search engines who used this technology to index his company's internal HR web site. He then measured the time it took him to process 15 typical HR requests, first using conventional search tools that he had available, and then using the Conceptual Indexing technology. In both cases, he measured the time it took him to either find the answer or to conclude that the answer wasn't in the indexed material. His measured times for the total suite were 55 minutes using the conventional

tools and 11 minutes using the conceptual indexing technology. Of course, this was an uncontrolled experiment, and there is some potential that information learned from searching with the traditional tools (which were apparently used first) might have provided some benefit when using the conceptual indexing technology. However, the fact that he found things with the latter that he did not find with the former and the magnitude of the time difference suggests that there is an effect, albeit perhaps not as great as the measurements. As a result of this experience, he concluded that he would expect many users to take much longer to find materials or give up, when using the traditional tools. He anticipated that after finding some initial materials, more time would be required, as users would end up having to call people for additional information. He estimated that users could spend up to an hour trying to get the information they needed...having to call someone, wait to make contact and finally get the information they needed. Using the conceptual indexing search engine, he expected that these times would be at least halved.

## 7 Conclusion

We have described some experiments using linguistic knowledge in an information retrieval system in which passages within texts are dynamically found in response to a query and are scored and ranked based on a relaxation of constraints. This is a different approach from previous methods of passage retrieval and from previous attempts to use linguistic knowledge in information retrieval. These experiments show that linguistic knowledge can significantly improve information retrieval performance when incorporated into a knowledge-based relaxation-ranking algorithm for specific passage retrieval.

The linguistic knowledge considered here includes the use of morphological relationships between words, taxonomic relationships between concepts, and general semantic entailment relationships between words and concepts. We have shown that the combination of these three knowledge sources can significantly improve performance in finding appropriate answers to specific queries when incorporated into a relaxation-ranking algorithm. It appears that the penalty-based relaxation-ranking algorithm figures crucially in this success, since the addition of such linguistic knowledge to traditional information retrieval models typically degrades retrieval performance rather than improving it, a pattern that was borne out in our own experiments.

## Acknowledgments

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



## References

Jacek Ambroziak and William A. Woods. 1998. Natural language technology in precision content retrieval. In *International Conference on Natural Language Processing and Industrial Applications*, Moncton, New Brunswick, Canada, August. www.sun.com/research/techrep/1998/abstract-69.html.

Jamie P. Callan. 1994. Passage-level evidence in document retrieval. *SIGIR*, pages 302–309.

J. L. Fagan. 1989. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2):115–132, March.

Marcin Kaszkiel and Justin Zobel. 1997. Passage retrieval revisited. *SIGIR*, pages 302–309.

David D. Lewis and Karen Sparck Jones. 1996. Natural language processing for information retrieval. *CACM*, 39(1):92–101.

Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM-SIGIR.

Gerald Salton, James Allan, and Chris Buckley. 1993. Approaches to passage retrieval in full text information systems. *SIGIR*, pages 49–58.

Gerard Salton. 1989. *Automatic Text Processing.* Addison Wesley, Reading, MA.

Karen Sparck Jones. 1998. A look back and a look forward. *SIGIR*, pages 13–29.

Giovanni Varile and Antonio Zampolli, editors. 1997. *Survey of the State of the Art in Human Language Technology.* Cambridge Univ. Press.

Ellen M. Voorhees. 1993. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of 16th ACM SIGIR Conference.* ACM-SIGIR.

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 *(these proceedings)*.