Conceptual Indexing: Practical Large-Scale AI for Efficient Information Access

William A. Woods
Sun Microsystems Laboratories
One Network Drive
Burlington, MA 01803
william.woods@east.sun.com

Abstract

Finding information is a problem shared by people and intelligent systems. This paper describes an experiment combining both human and machine aspects in a knowledge-based system to help people find information in text. Unlike many previous attempts, this system demonstrates a substantial improvement in search effectiveness by using linguistic and world knowledge and exploiting sophisticated knowledge representation techniques. It is also an example of practical subsumption technology on a large scale and with domain-independent knowledge. Results from this experiment are relevant to general problems of knowledge-based reasoning with large-scale knowledge bases.

Introduction

Long-term solutions to many problems in AI require efficient access to large amounts of knowledge. To be effectively scalable, such applications need to be able to find specific items of relevant information in an amount of time that is sublinear with respect to the size of the body of knowledge employed. This paper will describe the use of KL-One-style subsumption technology (Woods & Schmolze 1992) to address this problem in the context of a system to help people find specific information in online text. It will also discuss the extrapolation of these techniques to other large-scale knowledge-based applications.

The system in question 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). The system uses intensional subsumption technology to classify each word and phrase into a conceptual taxonomy in which each concept is linked to the most specific concepts that subsume it (i.e., that are more general). This taxonomy, together with records of where each concept occurs in the indexed material, constitutes the conceptual index. Subsumption paths through the taxonomy can be used to connect terms in an information request to related terms in indexed material, and the positional information in the index can be used to identify small, focused passages in the material where the requested information is likely to be.

This ability to locate specific answers to specific questions in unrestricted text is a useful intermediate capability between true question answering and traditional document retrieval. Experiments have shown that using this system results in a substantial improvement in human search productivity (Woods *et al.* 2000).

Background

I began this study by looking at what goes wrong when people fail to find what they want online. By catching people in what I came to call an information-seeking state, I was occasionally able to capture a spontaneous statement of the information need and watch the process by which people searched for information. In some cases I was able to compare what they eventually found to what they said they wanted. As a result of this study, I identified two principal problems that stand in the way of effective search. The first is that the terms used in the request are often different from the terms that are used by the author of the needed material (the paraphrase problem). The second is that it is usually difficult to find the relevant information in the documents that are retrieved (the information location problem).

I conjectured that an intensional subsumption methodology that I had been working on in the context of knowledge-based reasoning (Woods 1991) could be applied to both problems by locating specific places in the source material where particular things were said, and by dealing with the generality relationships that hold between the terms used in a query and those used in the indexed material. An experimental conceptual indexing and retrieval system was designed and constructed to test this hypothesis (Ambroziak & Woods 1998). Before describing it, I will first introduce the problems that the system needs to solve.

The problems

There are four sources of difficulty inherent in natural language that stand as potential obstacles to effective search. These problems are:

- Morphological variation – e.g., acid glass vs. acidic glass
- Differing terminology – e.g., moon rocks vs. lunar rocks
- Word relationships that alter meaning – e.g., man bites dog vs. dog bites man

- Word sense – e.g., dolphins as marine mammals vs. the Miami Dolphins

In the first case, morphological differences in inflection or in derivation give rise to different forms of a word. In the second, semantic relationships between different words with related meanings constitute the link between what was asked for and what was needed. In the third case, the mere presence of the requested elements is not sufficient to guarantee that they are being used in the desired relationships. In the fourth case, only some senses of ambiguous words are of interest.

To address these problems, we need to consider the following questions:

- What information is required to connect the concepts in a query to concepts in a relevant passage?
- How can this information be organized and used efficiently?
- To what extent can conceptual descriptions of the content of a document be automatically extracted and organized from the document itself?

Subsumption and paraphrase

In dealing with terminological variation, I chose to use an approach based on subsumption of less general terms by more general terms and subsumption of derived and inflected forms of words by their roots and base forms. This was chosen over the more traditional approach of using a synonym thesaurus and a stemming algorithm because it preserves information and gives control of the granularity of searching to the user. There are actually very few true synonyms in natural language, and if one puts terms that are not truly synonymous into a synonym thesaurus mechanism, the result is a lack of precision in being able to specify what is desired. In contrast, subsumption technology enables the searcher to specify a request at the level of generality of interest, and only those terms that are subsumed by the requested term are automatically searched for. Subsumption technology handles all of the phenomena of synonyms and more. In fact, synonymy is just the special case of mutual subsumption. Thus, in a technical sense, one can say that "subsumption subsumes synonymy."

With this approach, we answer the above questions in the following way:

- We use semantic relationships between concepts, morphological relationships between words, and syntactic relationships between words and phrases to connect terms of a request with relevant material.
- We use intensional subsumption algorithms to organize words and phrases into a conceptual taxonomy in order to use this information efficiently.
- We use a robust syntactic phrase extractor to extract words and phrases from the indexed material and parse them into conceptual structures which are then automatically organized into the conceptual taxonomy by taxonomic subsumption algorithms.

Conceptual Indexing

In conceptual indexing, the indexing system performs a significant, but limited amount of linguistic content processing at indexing time in order to support later retrieval operations. Every word that is encountered is looked up in a lexicon to determine its syntactic part of speech, any morphological structure that it exhibits, what specific word senses it may have, and any semantic relationships that it has to other words or word senses. Words that are not already in the lexicon are subjected to a knowledge-based morphological analysis (Woods 2000) in order to construct a new lexical entry that will be used for any subsequent occurrences of the same word. This system deals with prefixes, suffixes, and lexical compounds (e.g., *bitmap*). The system is capable of making plausible analyses of completely unknown words and is integrated with the conceptual taxonomy so that rules can make use of subsumption facts about words as well as syntactic and spelling information. Many rules also infer semantic subsumption relationships as a result of their analyses.

After lexical analysis, sequences of words that can form basic phrases (simple noun phrases and some simple verb phrases) are identified and parsed into conceptual structures. These structures are then automatically classified into the conceptual taxonomy by a taxonomic subsumption algorithm that can efficiently locate the place in the evolving taxonomy where the new concept should be placed. This algorithm, known as an MSS algorithm (for "most specific subsumer"), finds the most specific concepts in the taxonomy that subsume the new concept, and links the new concept directly under those concepts. A related algorithm, known as the MGS algorithm (for "most general subsumee"), finds the most general concepts that are subsumed by the new concept and links them directly under the new one.

When a single word is added to the taxonomy, any subsuming words or word senses that are listed in the lexical entry for that word are also added to the taxonomy and subsumption links are created to link these words and word senses to their most specific subsumers. When a phrase is added, the MSS algorithms searches the conceptual taxonomy to find its most specific subsumers, using the conceptual structures of the phrases to align corresponding constituent elements and using semantic subsumption relationships already in the taxonomy to relate constituents to each other. For example, in one taxonomy, the phrase *automobile cleaning* subsumes *car washing* because of the way corresponding elements of these phrases are aligned and the facts (derived from semantic subsumption information in lexical entries) that a car is a kind of automobile and washing is a kind of cleaning.

Because of the conceptual subsumption algorithms, it is possible to look up a phrase that doesn't occur in the taxonomy and still find useful concepts in the "neighborhood." This is because the algorithm automatically locates the places in the taxonomy where the query concept would belong, using the same MSS and MGS algorithms that are used to construct the taxonomy. For example, one can look up *automobile cleaning* in a taxonomy and find *car washing.* One can think of the subsumption logic that governs the taxonomy as providing a conceptual space in which to search that is an "oriented topology." The orienting principle is the up/down relationship of generality and the topological neighborhoods are determined (and linguistically labeled) by the subsuming concepts and their descendants. This conceptual space provides a naturally intuitive structure for browsing and navigating, and allows subsumption algorithms to be used to get quickly to the right neighborhood (without having to navigate all the way down from the top). This is a more effective structure to navigate than a strictly hierarchical classification tree, and it provides more informative browsing capabilities than approaches based on linking concepts according to measures of term-term distance.

The utility of a conceptual index can be illustrated by an example (shown in Figure 1) from the conceptual index of a collection of encyclopedia articles about animals. An initial request for *brown fur* retrieved the phrase `(BROWN FUR)` and the subsumed phrases `(GRAY BROWN FUR)`, `(RICH BROWN FUR)`, and `(WHITE-SPOTTED BROWN FUR)`. However, a display of more general concepts showed that the query was subsumed by the phrase `(BROWN COAT)`, revealing that the request was inadvertently more specific than intended. Generalizing the request to `(BROWN COAT)` produced the substantially more useful collection of concepts shown in Figure 1. By displaying the more general concepts in the taxonomy that subsumed the stated request, the system unobtrusively suggested an exceedingly useful generalization of the request.

Note that morphological relationships are incorporated into the subsumption framework by treating derived and inflected forms of words as subsumed by their base forms. For example, *brownish* is subsumed by *brown.* Morphological variations and terminology variations are thus automatically related in the conceptual taxonomy, and syntactic relationships are incorporated into the structures of the parsed phrases. Different senses of words can be represented by distinct concepts that have different places in the taxonomy but are subsumed by an abstract concept corresponding to the undisambiguated word.

Specific Passage Retrieval

The conceptual index supports direct access to places in the indexed material where concepts subsumed by a query occur. However, there are often situations in which elements of a query occur in a text without being explicitly related in ways that would enable a strict subsumption. For example, in the Encyclopedia example illustrated above, subsumption fails to pick up a passage of text that says, *The coat is reddish brown,* because there is no explicit phrase subsumed by *brown coat.* To handle such cases, we use a technique called "relaxation ranking" (Woods 1997; Ambroziak & Woods 1998; Woods *et al.* 2000) to find specific passages where as many as possible of the elements of a query occur near each other, preferably in the same form and word order and preferably closer together. These passages are given a penalty score based on the amount of deviation from an exact copy of the query (same words, same forms, same order, and nothing else in between). Penalties are assigned in proportion to the amount of intervening material occurring between the desired terms. When the elements occur in a different order for reasons that are not accounted for, an extra penalty is assigned proportional to the amount of reordering of terms. In addition, small penalties can be assigned for subsumed terms and different inflected forms of the requested terms, and substantial penalties are assigned for elements that are missing from passages that contain some, but not all, of the elements of a request.

```
(BROWN COAT)
  |-k- (BRIGHT REDDISH BROWN COAT)
  |-k- (BROWN BLACK COAT)
  |-k- (BROWN COATS)
  |     |-k- (FAWN COATS)
  |     |     |-v- ((FAWN) COATS)
  |     |
  |     |-k- (REDDISH BROWN UPPER COATS)
  |
  |-k- (BROWN FUR)
  |     |-k- (GRAY BROWN FUR)
  |     |-k- (RICH BROWN FUR)
  |     |-k- (WHITE-SPOTTED BROWN FUR)
  |
  |-k- (BROWN HAIR)
  |     |-k- (BROWN HAIRS)
  |     |     |-k- (REDDISH BROWN GUARD HAIRS)
  |     |
  |     |-k- (BROWN WOOL)
  |     |     |-k- (REDDISH BROWN WOOL)
  |     |
  |     |-k- (BROWNISH HAIR)
  |     |-k- (REDDISH BROWN HAIR)
  |           |-k- (REDDISH BROWN GUARD HAIRS)
  |           |-k- (REDDISH BROWN WOOL)
  |
  |-k- (BROWN-GRAY COAT)
  |-k- (BROWNISH COAT)
  |     |-k- (BROWNISH HAIR)
  |     |-k- (BROWNISH SUMMER COAT)
  |     |-k- (COARSER BROWNISH COAT)
  |     |-k- (DARK BROWNISH COAT)
  |
  |-k- (COAT OF BROWN)
  |     |-k- (WOOLLY COAT OF REDDISH BROWN)
  |-k- (TAWNY COAT)
```

Figure 1: A fragment of a conceptual taxonomy, illustrating the utility of navigating in conceptual space.

Retrieved passages are constructed dynamically in response to a query, using information from the conceptual index about where concepts occur in the text. The hit passages from a single query may range in size from a single word or phrase to several sentences or a paragraph, with penalty scores generally increasing as passages get longer. This is in contrast with passage retrieval methods in which the indexed material is presegmented into paragraphs or sentences at indexing time. Because of the way scores are assigned, the penalty scores are highly correlated with the likelihood that the terms in the passage are related in the way that is requested, so the best passages tend to occur at the top of the ranking. This is because the likelihood that the terms are related in the desired way and are being used in the desired senses tends to decrease as the terms in the passage get further apart and their order departs from that of the input request.

For example, the following is a passage retrieved by this system, when applied to the UNIX® operating system on-line documentation (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 words in the input query. In this case, *print* is matched to *print*, *message* to *mail*, *mail* to *mail*, and *tool* to *mailtool*. 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*.

Precision Search Benefits

Precision Content Retrieval offers users three key benefits. The first is specific passage retrieval–that is, finding specific passages of information that are responsive to specific requests. The second is an effective ranking and scoring of the found passages, so that the best passages are listed first. The third is "conceptual navigation," the ability to move around in the structured conceptual taxonomy, a "conceptual space" that is intuitively organized for efficient browsing and navigation and includes all of the concepts found in the indexed material.

In quantitative evaluations of these benefits (Woods *et al.* 2000), one experiment showed that without any of its knowledge sources, the relaxation ranking algorithm is roughly as good as a state-of-the art commercial search engine at finding documents, and it becomes substantially better (41% better) when semantic and morphological knowledge is used. This is in contrast to most previous attempts to use linguistic knowledge to improve information retrieval (Fagan 1989; Lewis & Sparck Jones 1996; Mauldin 1991; Sparck Jones 1998; Varile & Zampolli 1997; Voorhees 1993; Mandala, Tokunaga, & Tanaka 1999) (but see the latter for some successes as well). This experiment doesn't measure the benefit of being able to locate the specific passages where the information occurs.

In another experiment, informally measuring the time it took to find specific information on a web site or to conclude that it is not in the indexed material, using a commercial search engine versus our conceptual indexing technology, showed a five-fold improvement in search productivity from using the conceptual indexing technology (55 minutes versus 11 minutes to work through a suite of 15 typical queries).

Large-Scale Knowledge Representation

The system that I have just described is an example of a large-scale application of subsumption technology. When we began this project, it was not clear that it would be possible to build large conceptual taxonomies in a reasonable amount of time. In fact, most of the theoretical work on KL-One-style techniques had produced unattractive complexity results at that time (Woods & Schmolze 1992). However, (Woods 1991) described an approach, called "intensional subsumption," that appeared likely to resolve some of the complexity issues and would at the same time increase the expressive power of the representational system. The system I have just described is the first large-scale test of this approach, and so far, the predictions of (Woods 1991) seem to have held up.

To date, the largest conceptual index that we have built contained over three million concepts. It was derived from a 3.1-million-word corpus of technical material. The core lexicon that we used contains approximately 80,000 words (150,000 word forms) and contains semantic subsumption facts for more than 15,000 words. More than 18,000 words were encountered in the 3.1-million-word corpus that were not in this lexicon (not counting numbers and hyphenated words), illustrating the importance of morphological analysis. The morphological analysis system that we used to analyze these words (and the hyphenated words as well) contains approximately 1,200 morphological rules. In addition to patterns of spelling of the analyzed words, these rules test syntactic and sometimes semantic information about hypothesized root words, in order to form hypotheses about the structure and meanings of the unknown words and to construct lexical entries for them.

When I began the exploration of subsumption technology for conceptually indexing text, one of the things I was interested in understanding was the behavior of taxonomic classification algorithms on large collections of naturally occurring concepts. The automatic extraction of words and phrases from unrestricted text provides an almost unlimited supply of naturally occurring concepts. This provides a more realistic testing environment than would an algorithmic study on synthesized examples. While we are still in the process of studying the behavior of large-scale classification algorithms, we have at least demonstrated that large scale conceptual taxonomies can be constructed and used and that they provide real benefit for the information location problem.

Other Applications of Large-Scale Subsumption

The intensional subsumption technology described here was originally motivated by the problem of organizing the rules of a large-scale knowledge-based reasoning system that would be able to deal with millions of rules and handle the kinds of fluent shifts in levels of generality that people exhibit when reasoning (Woods 1986). The idea was that the pattern parts of rules could be organized into a subsumption taxonomy so that the most specific rules that match a given subgoal could be found efficiently using the MSS algorithm. Although there are likely to be some differences between classifying the pattern parts of rules and the classification of natural English phrases, our experiences with the latter nevertheless give some idea of the feasibility and complexity of such classifications in large taxonomies. Moreover, there are reasons to suspect that the behavior for rule patterns might not be too different. For example, predicates used in logical rules are usually amenable to translation into ordinary English.

Our experiences with classifying English phrases show that it is at least possible to construct and use large conceptual taxonomies and to deal with the subsumption relationships between different levels of generality. It would be interesting to try to extend this approach to handling large-scale rule-based systems. Woods (1991) presents a number of advantages of such an approach, such as the ability to quickly determine the most specific subsuming rule patterns from very large collections of rules. In addition to the benefit of finding matching rules efficiently from a large collection, this approach has the added benefit of automatically organizing the pattern parts of the rules by generality, so that the most specific matching patterns are found, and any information from more general rules can be found by inheritance or overridden by the more specific rules.

Woods (1991) also presents an analysis of the structure of concepts in representations such as frames, semantic networks, object-oriented class systems, and conventional data base records, that makes it possible to correctly apply subsumption logic to such representations. One of the key insights is the observation that slots in frames, fields in data records, and class and instance variables in object-oriented systems, all have a quantificational import that determines how those elements participate in inheritance and subsumption reasoning. By making this quantificational import explicit, in the form of a system of quantificational tags, it is possible to develop a clean separation between the logical structure of concepts, which can be interpreted by subsumption and inheritance algorithms, and the domain-specific content of those concepts, which is then reasoned about by the subsumption logic and reasoning system. For example, different quantificational tags are used to express the different relationships implied by the superficially similar statements:

- birds have wings
- people need vitamins

In the first case, for every bird, there are some wings that it has, while in the second, for every person and every vitamin, that person needs that vitamin.

Quantificational tags can also be used to distinguish relationships that are governed by statistical probabilities, as opposed to strict logical implications, and those that are definitional versus those that are assertional. Using these distinctions, a subsumption taxonomy can integrate probabilistic with strictly logical information, as well as absolute and defeasible inheritance.

Furthermore, a classification process based on intensional subsumption can be used to efficiently resolve conflicts between competing rules in multiple inheritance situations. For example, in the famous "Nixon diamond" scenario, in which Richard Nixon, who was both a Republican and a Quaker, would be expected to inherit hawkishness from the `REPUBLICAN` concept and non-hawkishness from the `QUAKER` concept, a conceptual subsumption approach can identify the locus of the problem at the level of the greatest-lower-bound concept, `REPUBLICAN QUAKER`. This then suggests that an experiment needs to be conducted to determine the likelihoods of hawkishness for Republican Quakers. (This is an empirical fact that needs to be determined, not something that can be resolved by any principle of inheritance.) Once such likelihoods have been determined and associated with the `REPUBLICAN QUAKER` concept, individual instances like Richard Nixon, that would be classified under both `REPUBLICAN` and `QUAKER`, will automatically be classified under this combined concept by the MSS algorithm. Thus, they will inherit the resolved result from the immediate parent concept, `REPUBLICAN QUAKER`, without needing to even notice the conflicting sources of information above it.

The detection of such potential conflicts can be done either as a consequence of finding an individual case which inherits conflicting information (in which case a combined concept is then created from the parents that participate in the conflict), or it can be done as a background process, searching for greatest-lower-bound concepts whose parent concepts would provide conflicting information. In the latter case, one need not wait to encounter conflicting instances in order to ask the necessary questions to resolve the conflicts.

In summary, the organizing principles of intensional subsumption hold a lot of promise for organizing very large collections of rules for efficient use. The experiments in conceptual indexing of text suggest that such an approach might be tractable. However, it is difficult to test this hypothesis without a problem that provides a source of millions of rules. One wonders what a naturally occurring population of such rule patterns might be like and how a taxonomic classification approach would behave on such a population.

Conclusion

I have described a system that constructs and uses large-scale conceptual taxonomies of English concepts that are automatically extracted from unrestricted text. These taxonomies of meaningful words and phrases are used to help people find specific information in online text by efficiently finding connections between terms used in a request and related terms used by authors of the text material. It provides an example of practical subsumption technology on a large scale and with domain-independent knowledge.

This system has demonstrated a substantial improvement in search effectiveness from using linguistic and world knowledge and exploiting intensional subsumption technology in conjunction with a relaxation-ranking passage retrieval algorithm. Conceptual taxonomies in excess of three million concepts have been automatically constructed, indexing previously unseen bodies of text material and using a pre-existing body of general-purpose linguistic and semantic rules and knowledge.

Results from this experiment are relevant to general problems of knowledge-based reasoning with large-scale knowledge bases and with domain-independent knowledge. An important issue is the problem of scale – how to find a small amount of information relevant to a given problem out of a truly encyclopedic body of knowledge. Intensional subsumption technology (Woods 1991) is used to find the most specific subsumers of a concept in an amount of time that is sublinear in the size of the taxonomy. Such taxonomies can be used to organize the pattern parts of rules, as well as facts, and thus find matching rules from very large rule sets in a reasonable amount of time. Subsumption technology can also efficiently resolve conflicting inheritance paths.

The conceptual indexing system described is an example of practical AI using large-scale and domain-independent knowledge bases that provides a significant improvement in human information access.

Acknowledgments

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

References

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

Fagan, J. L. 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.

Lehmann, F., ed. 1992. *Semantic Networks in Artificial Intelligence*. Pergamon Press.

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

Mandala, R.; Tokunaga, T.; and Tanaka, H. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM-SIGIR.

Mauldin, M. L. 1991. Retrieval performance in FERRET a conceptual information retrieval system. In *14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 347–355.

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

Varile, G., and Zampolli, A., eds. 1997. *Survey of the State of the Art in Human Language Technology*. Cambridge Univ. Press.

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

Woods, W. A., and Schmolze, J. G. 1992. The kl-one family. *Computers & Mathematics with Applications* 23(2–5):133–177. Reprinted in (Lehmann, 1992).

Woods, W. A.; Bookman, L. A.; Houston, A. C.; Kuhns, R. J.; Martin, P. A.; and Green, S. 2000. Linguistic knowledge can improve information retrieval. In *Proceedings of ANLP-2000*. Seattle WA: ACL ANLP.

Woods, W. A. 1986. Important issues in knowledge representation. *Proceedings of the IEEE* 74(10):1322–1334. Reprinted in (Woods, 1990).

Woods, W. A. 1990. Important issues in knowledge representation. In Raeth, P. G., ed., *Expert Systems: A Software Methodology for Modern Applications*. IEEE Computer Society Press. 180–204.

Woods, W. A. 1991. Understanding subsumption and taxonomy: A framework for progress. In Sowa, J., ed., *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann. 45–94.

Woods, W. A. 1997. Conceptual indexing: A better way to organize knowledge. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA. www.sun.com/research/techrep/1997/abstract-61.html.

Woods, W. A. 2000. Aggressive morphology for robust lexical coverage. In *Proceedings of ANLP-2000*. Seattle WA: ACL ANLP.