quantitative information is available about the extent of the problem or about the impact that it has on information retrieval systems. We report on an analysis of lexical ambiguity in information retrieval test collections and on experiments to determine the utility of word meanings for separating relevant from nonrelevant documents. The experiments show that there is considerable ambiguity even in a specialized database. Word senses provide a significant separation between relevant and nonrelevant documents, but several factors contribute to determining whether disambiguation will make an improvement in performance For example, resolving lexical ambiguity was found to have little impact on retrieval effectiveness for documents that have many words in common with the query. Other uses of word sense disambiguation in an information retrieval context are discussed

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*dictionaries, indexing methods, linguistic processing;* H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*search process, selection process;* I.2.7 [Artificial Intelligence]: Natural Language Processing—*text analysis*

General Terms: Experimentation, Measurement, Performance

Additional Key Words and Phrases: Disambiguation, document retrieval, semantically based search, word senses

1. INTRODUCTION

The goal of an information retrieval system is to locate relevant documents in response to a user's query. Documents are typically retrieved as a ranked list, where the ranking is based on estimations of relevance [5]. The *retrieval model* for an information retrieval system specifies how documents and queries are represented and how these representations are compared to produce relevance estimates. The performance of the system is evaluated

This work has been supported by the Office of Naval Research under University Research Initiative Grant N00014-86-K-0746, by the Air Force Office of Scientific Research, under contract 91-0324, and by NSF Grant IRI-8814790.

Authors' address: Computer Science Department, University of Massachusetts, Amherst, MA 01003; email:krovetz@cs.umass.edu and croft@cs.umass.edu.

Performance is improved by weighting query and document words using frequency information from the collection and individual document texts [27].

There are two problems with using words to represent the content of documents. The first problem is that words are ambiguous, and this ambiguity can cause documents to be retrieved that are not relevant. Consider the following description of a search that was performed using the keyword "AIDS":

> Unfortunately, not all 34 [references] were about AIDS, the disease. The references included "two helpful aids during the first three months after total hip replacement," and "aids in diagnosing abnormal voiding patterns" [17].

One response to this problem is to use phrases to reduce ambiguity (e.g., specifying "hearing aids" if that is the desired sense) [27]. It is not always possible, however, to provide phrases in which the word occurs only with the desired sense. In addition, the requirement for phrases imposes a significant burden on the user.

The second problem is that a document can be relevant even though it does not use the same words as those that are provided in the query. The user is generally not interested in retrieving documents with exactly the same words, but with the concepts that those words represent. Retrieval systems address this problem by expanding the query words using related words from a thesaurus [27]. The relationships described in a thesaurus, however, are really between word senses rather than words. For example, the word "term" could be synonymous with "word" (as in a vocabulary term), "sentence" (as in a prison term), or "condition" (as in "terms of agreement"). If we expand the query with words from a thesaurus, we must be careful to use the right senses of those words. We not only have to know the sense of the word in the query (in this example, the sense of the word "term"), but the sense of the word that is being used to augment it (e.g., the appropriate sense of the word "sentence") [7].[1]

[1] Salton recommends that a thesaurus should be coded for ambiguous words, but only for those senses likely to appear in the collections to be treated [26, pp. 28–29]. However, it is not always easy to make such judgments, and it makes the retrieval system specific to particular subject areas. The thesauri that are currently used in retrieval systems do not take word senses into account.

In these experiments, word senses are taken from a machine-readable dictionary. Dictionaries vary widely in the information they contain and the number of senses they describe. At one extreme we have pocket dictionaries with about 35,000–45,000 senses, and at the other the *Oxford English Dictionary*, with over 500,000 senses and in which a single entry can go on for several pages. Even large dictionaries will not contain an exhaustive listing of all of a word's senses; a word can be used in a technical sense specific to a particular field, and new words are constantly entering the language. It is important, however, that the dictionary contain a variety of information that can be used to distinguish the word senses. The dictionary we are using in our research, the *Longman Dictionary of Contemporary English* (LDOCE) [25], has the following information associated with its senses: part of speech, subcategorization,[2] morphology, semantic restrictions, and subject classification.[3] The latter two are only present in the machine-readable version.

In the following section we discuss previous research that has been done on lexical ambiguity and its relevance to information retrieval. This includes work on the types of ambiguity and algorithms for word sense disambiguation. In Section 3 we present and analyze the results of a series of experiments on lexical ambiguity in information retrieval test collections.

2. PREVIOUS RESEARCH ON LEXICAL AMBIGUITY

2.1 Types of Lexical Ambiguity

The literature generally divides lexical ambiguity into two types: syntactic and semantic [31]. Syntactic ambiguity refers to differences in syntactic category (e.g., *play* can occur as either a noun or verb). Semantic ambiguity refers to differences in meaning, and is further broken down into homonymy or polysemy, depending on whether or not the meanings are related. The *bark* of a dog versus the *bark* of a tree is an example of homonymy; *opening* a door versus *opening* a book is an example of polysemy. Syntactic and

[2] This refers to subclasses of grammatical categories such as transitive versus intransitive verbs.
[3] Not all senses have all of this information associated with them. Also, some information, such as part of speech and morphology, is associated with the overall headword rather than just the sense.

It also is not clear how the relationship of senses affects their role in information retrieval. Although senses which are unrelated might be more useful for separating relevant from nonrelevant documents, we found a number of instances in which related senses also acted as good discriminators (e.g., "West Germany" versus "The West").

2.2 Automatic Disambiguation

A number of approaches have been taken to word sense disambiguation. Small used a procedural approach in the Word Experts system [30]: words are considered experts of their own meaning and resolve their senses by passing messages between themselves. Cottrell resolved senses using connectionism [9], and Hirst and Hayes made use of spreading activation and semantic networks [18, 16].

Perhaps the greatest difficulty encountered by previous work was the effort required to construct a representation of the senses. Because of the effort required, most systems have only dealt with a small number of words and a subset of their senses. Small's Word Expert Parser only contained Word Experts for a few dozen words, and Hayes' work only focused on disambiguating nouns. Another shortcoming is that very little work has been done on disambiguating large collections of real-world text. Researchers have instead argued for the advantages of their systems based on theoretical grounds and shown how they work over a selected set of examples. Although information retrieval test collections are small compared to real world databases, they are still orders of magnitude larger than single sentence examples. Machine-readable dictionaries give us a way to temporarily avoid the problem of representation of senses.[4] Instead the work can focus on how well information about the occurrence of a word in context matches with the information associated with its senses.

It is currently not clear what kinds of information will prove most useful for disambiguation. In particular, it is not clear what kinds of knowledge will be required that is not contained in a dictionary. In the sentence "John left

[4] We will eventually have to deal with word sense representation because of problems associated with dictionaries being incomplete, and because they may make too *many* distinctions; these are important research issues in lexical semantics. For more discussion on this see Krovetz [21].

by Weiss in the context of information retrieval [34]. Words are disambiguated via two kinds of rules: template rules and contextual rules. There is one set of rules for each word to be disambiguated. Template rules look at the words that cooccur within two words of the word to be disambiguated; contextual rules allow a range of five words and ignore a subset of the closed-class words (words such as determiners, prepositions and conjunctions). In addition, template rules are ordered before contextual rules. Within each class, rules are manually ordered by their frequency of success at determining the correct sense of the ambiguous word. A word is disambiguated by trying each rule in the rule set for the word, starting with the first rule in the set and continuing with each rule in turn until the cooccurrence specified by the rule is satisfied. For example, the word "type" has a rule that indicates if it is followed by the word "of" then it has the meaning "kind" (a template rule); if "type" cooccurs within five words of the word "pica" or "print," it is given a printing interpretation (a contextual rule). Weiss conducted two sets of experiments: one on five words that occurred in the queries of a test collection on documentation and one on three words, but with a version of the system that learned the rules. Weiss felt that disambiguation would be more useful for question answering than strict information retrieval, but would become more necessary as databases became larger and more general.

Word collocation was also used in several other disambiguation efforts. Black compared collocation with an approach based on subject-area codes and found collocation to be more effective [6]. Dahlgren used collocation as one component of a multiphase disambiguation system (she also used syntax and "common sense knowledge" based on the results of psycholinguistic studies) [12]. Atkins examined the reliability of collocation and syntax for identifying the senses of the word "danger" in a large corpus [3]; she found that they were reliable indicators of a particular sense for approximately 70% of the word instances she examined. Finally, Choueka and Lusignan showed that people can often disambiguate words with only a few words of context (frequently only one word is needed) [8].

Syntax is also an important source of information for disambiguation. Along with the work of Dahlgren and Atkins, it has also been used by Kelly and Stone for content analysis in the social sciences [20], and by Earl for machine translation [13]. The latter work was primarily concerned with subcategorization (distinctions within a syntactic category), but also included

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.