Lexical Ambiguity and Information Retrieval

Robert Krovetz
W. Bruce Croft
Computer and Information Science Department
University of Massachusetts, Amherst, MA 01003

Abstract

Lexical ambiguity is a pervasive problem in natural language processing. However, little 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 non-relevant documents. The experiments show that there is considerable ambiguity even in a specialized database. Word senses provide a significant separation between relevant and non-relevant 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: Word senses, disambiguation, document retrieval, semantically based search

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 with respect to standard test collections that provide a set of queries, a set of documents, and a set of relevance judgments that indicate which documents are relevant to each query. These judgments are provided by the users who supply the queries, and serve as a standard for evaluating performance. Information retrieval research is concerned with finding representations and methods of comparison that will accurately discriminate between relevant and non-relevant documents.

Many retrieval systems represent documents and queries by the words they contain, and base the comparison on the number of words they have in common. The more words the query and document have in common, the higher the document is ranked; this is referred to as a 'coordination match'. 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]

It is possible that representing documents by word senses, rather than words, will improve retrieval performance. Word senses represent more of the semantics of the text, and they provide a basis for exploring lexical semantic relationships such as synonymy and antonymy, which are important in the construction of thesauri. Very little is known, however, about the quantitative aspects of lexical ambiguity. In this paper, we describe experiments designed to discover the degree of lexical ambiguity in information retrieval test collections, and the utility of word senses for discriminating between relevant and non-relevant documents. The data from these experiments will also provide guidance in the design of algorithms for automatic disambiguation.

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.

[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.

[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.

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 a 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 semantic ambiguity are orthogonal, since a word can have related meanings in different categories ('He will *review* the *review* when he gets back from vacation'), or unrelated meanings in different categories ('*Can* you see the *can*?').

Although there is a theoretical distinction between homonomy and polysemy, it is not always easy to tell them apart in practice. What determines whether the senses are related? Dictionaries group senses based on part-of-speech and etymology, but as mentioned above, senses can be related even though they differ in syntactic category. Senses may also be related etymologically, but be perceived as distinct at the present time (e.g., the 'cardinal' of a church and 'cardinal' numbers are etymologically related). 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 non-relevant 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 are not contained in a dictionary. In the sentence 'John left a tip', the word 'tip' might mean a gratuity or a piece of advice. Cullingford and Pazzani cite this as an example in which scripts are needed for disambiguation [11]. There is little data, however, about how often such a case occurs, how many scripts would be involved, or how much effort is required to construct them. We might be able to do just as well via the use of word co-occurrences (the gratuity sense of tip is likely to occur in the same context as 'restaurant', 'waiter', 'menu', etc.). That is, we might be able to use the words that could trigger a script without actually making use of one.

Word co-occurrences are a very effective source of information for resolving ambiguity, as will be shown by experiments described in section 3. They also form the basis for one of the earliest disambiguation systems, which was developed 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 co-occur 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, conjunctions, etc.). 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 co-occurrence 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' co-occurs 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,

[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 [21].

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 multi-phase 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 semantic categories as part of the patterns associated with various words. Earl and her colleagues noticed that the patterns could be used for disambiguation, and speculated that they might be used in information retrieval to help determine better phrases for indexing.

Finally, the redundancy in a text can be a useful source of information. The words 'bat', 'ball', 'pitcher', and 'base' are all ambiguous and can be used in a variety of contexts, but collectively they indicate a single context and particular meanings. These ideas have been discussed in the literature for a long time ([2], [24]) but have only recently been exploited in computerized systems. All of the efforts rely on the use of a thesaurus, either explicitly, as in the work of Bradley and Liaw (cf. [28]), or implicitly, as in the work of Slator [29]. The basic idea is to compute a histogram over the classes of a thesaurus; for each word in a document, a counter is incremented for each thesaurus class in which the word is a member. The top rated thesaurus classes are then used to provide a bias for which senses of the words are correct. Bradley and Liaw use Roget's Third International Thesaurus, and Slator uses the subject codes associated with senses in the Longman Dictionary of Contemporary English (LDOCE).[5]

Machine readable dictionaries have also been used in two other disambiguation systems. Lesk, using the Oxford Advanced Learners Dictionary,[6] takes a simple approach to disambiguation: words are disambiguated by counting the overlap between words used in the

[5] These codes are only present in the machine readable version.

[6] Lesk also tried the same experiments with the Merriam-Webster Collegiate Dictionary and the Collins English Dictionary; while he did not find any significant differences, he speculated that the longer definitions used in the Oxford English Dictionary (OED) might yield better results. Later work by Becker on the New OED indicated that Lesk's algorithm did not perform as well as expected [4].

definitions of the senses [23]. For example, the word 'pine' can have two senses: a tree, or sadness (as in 'pine away'), and the word 'cone' may be a geometric structure, or a fruit of a tree. Lesk's program computes the overlap between the senses of 'pine' and 'cone', and finds that the senses meaning 'tree' and 'fruit of a tree' have the most words in common. Lesk gives a success rate of fifty to seventy percent in disambiguating the words over a small collection of text.

Wilks performed a similar experiment using the Longman dictionary [35]. Rather than just counting the overlap of words, all the words in the definition of a particular sense of some word are grouped into a vector. To determine the sense of a word in a sentence, a vector of words from the sentence is compared to the vectors constructed from the sense definitions. The word is assigned the sense corresponding to the most similar vector. Wilks manually disambiguated all occurrences of the word 'bank' within LDOCE according to the senses of its definition and compared this to the results of the vector matching. Of the 197 occurrences of 'bank', the similarity match correctly assigned 45 percent of them to the correct sense; the correct sense was in the top three senses 85 percent of the time.

Because information retrieval systems handle large text databases (megabytes for a test collection, and gigabytes/terabytes for an operational system), the correct sense will never be known for most of the words encountered. This is due to the simple fact that no human being will ever provide such confirmation. In addition, it is not always clear just what the 'correct sense' is. In disambiguating the occurrences of 'bank' within the Longman dictionary, Wilks found a number of cases where none of the senses was clearly 'the right one' [35]. In the information retrieval context, however, it may not be necessary to identify the single correct sense of a word; retrieval effectiveness may be improved by ruling out as many of the incorrect word senses as possible, and giving a high weight to the senses most likely to be correct.

Another factor to consider is that the dictionary may sometimes make distinctions that are not necessarily useful for a particular application. For example, consider the senses for the word 'term' in the Longman dictionary. Seven of the senses are for a noun, and one is for a verb. Of the seven noun senses, five refer to periods of time; one has the meaning 'a vocabulary item'; and one has the meaning 'a component of a mathematical expression'. It may only be important to distinguish the four classes (three noun and one verb), with the five 'period of time' senses being collapsed into one. The experiments in this paper provide some insight into the important sense distinctions for information retrieval.

As we mentioned at the start of this section, a major problem with previous approaches has been the effort required to develop a lexicon. Dahlgren is currently conducting tests on a 6,000 word corpus based on six articles from the Wall Street Journal. Development of the lexicon (which includes entries for 5,000 words)[7] took 8 man-years of effort (Dahlgren, personal communication). This effort did not include a representation for all of the senses for those words, only the senses that actually occurred in the corpora she has been studying. While a significant part of this time was devoted to a one-time design effort, a substantial amount of time is still required for adding new words.

The research described above has not provided many experimental results. Several researchers did not provide any experimental evidence, and the rest only conducted experiments on a small collection of text, a small number of words, and/or a restricted range of senses. Although some work has been done with information retrieval collections (e.g., [34]), disambiguation was only done for the queries. None of the previous work has provided evidence that disambiguation would be useful in separating relevant from non-relevant documents. The following sections will describe the degree of ambiguity found in two information retrieval test collections, and experiments involving word sense weighting, word sense matching, and the distribution of senses in queries and in the corpora.

3 Experimental Results on Lexical Ambiguity

Although lexical ambiguity is often mentioned in the information retrieval literature as a problem (cf. [19], [26]), relatively little information is provided about the degree of ambiguity encountered, or how much improvement would result from its resolution.[8] We conducted experiments to determine the effectiveness of weighting words by the number of senses they have, and to determine the utility of word meanings in separating relevant from non-relevant documents. We will first provide statistics about the retrieval collections we used, and then describe the results of our experiments.

3.1 Collection Statistics

Information retrieval systems are evaluated with respect to standard test collections. Our experiments were done on two of these collections: a set of titles and abstracts from Communications of the ACM (CACM) [14] and a set of short articles from TIME magazine. We chose these collections because of the contrast they provide; we wanted to see whether the subject area of the text has any effect on our experiments. Each collection also includes a set

[7] These entries are based not only on the Wall Street Journal corpus, but a corpus of 4100 words taken from a geography text.

[8] Weiss mentions that resolving ambiguity in the SMART system was found to improve performance by only 1 percent, but did not provide any details on the experiments that were involved [34].

	CACM	TIME
Number of queries	64	83
Number of documents	3204	423
Mean words per query	9.46	7.44
Mean words per document	94	581
Mean relevant documents per query	15.84	3.90

Table 1: Statistics on information retrieval test collections

of natural language queries and relevance judgments that indicate which documents are relevant to each query. The CACM collection contains 3204 titles and abstracts[9] and 64 queries. The TIME collection contains only 423 documents[10] and 83 queries, but the documents are more than six times longer than the CACM abstracts so the collection overall contains more text. Table 1 lists the basic statistics for the two collections. We note that there are far fewer relevant documents per query for the TIME collection than for the CACM collection. The average for CACM does not include the 12 queries that do not have relevant documents.

Table 2 provides statistics about the word senses found in the two collections. The mean number of senses for the documents and queries was determined by a dictionary lookup process. Each word was initially retrieved from the dictionary directly; if it was not found the lookup was retried, this time making use of a simple morphological analyzer.[11] For each dataset, the mean number of senses is calculated by averaging the number of senses for all unique words (word types) found in the dictionary.

The statistics indicate that a similar percentage of the words in the TIME and CACM collections appear in the dictionary (about 40% before any morphology, and 57 to 65% once simple morphology is done),[12] but that the TIME collection contains about twice as many unique words as CACM. Our morphological analyzer primarily does inflectional morphology (tense, aspect, plural, negation, comparative, and superlative). We estimate that adding more

[9] Half of these are title only.

[10] The original collection contained 425 documents, but two of the documents were duplicates.

[11] This analyzer is not the same as a 'stemmer', which conflates word variants by truncating their endings; a stemmer does not indicate a word's root, and would not provide us with a way to determine which words were found in the dictionary. Stemming is commonly used in information retrieval systems, however, and was therefore used in the experiments that follow.

[12] These percentages refer to the unique words (word types) in the corpora. The words that were not in the dictionary consist of hyphenated forms, proper nouns, morphological variants not captured by the simple analyzer, and words that are domain specific.

CACM

	Unique Words	Word Occurrences
Number of words in the corpus	10203	169769
Number of those words in LDOCE	3922 (38%)	131804 (78%)
Including morphological variants	5799 (57%)	149358 (88%)
Mean number of senses in the collection	4.7 (4.4 without stop words)	
Mean number of senses in the queries	6.8 (5.3 without stop words)	

TIME

	Unique Words	Word Occurrences
Number of words in the corpus	22106	247031
Number of those words in LDOCE	9355 (42%)	196083 (79%)
Including morphological variants	14326 (65%)	215967 (87%)
Mean number of senses in the collection	3.7 (3.6 without stop words)	
Mean number of senses in the queries	8.2 (4.8 without stop words)	

Table 2: Statistics for word senses in IR test collections

complex morphology would capture another 10 percent.

The statistics indicate that both collections have the potential to benefit from disambiguation. The mean number of senses for the CACM collection is 4.7 (4.4 once stop words are removed)[13] and 3.7 senses for the TIME collection (3.6 senses without the stop words). The ambiguity of the words in the queries is also important. If those words were unambiguous then disambiguation would not be needed because the documents would be retrieved based on the senses of the words in the queries. Our results indicate that the words in the queries are even more ambiguous than those in the documents.

[13] Stop words are words that are not considered useful for indexing, such as determiners, prepositions, conjunctions, and other closed class words. They are among the most ambiguous words in the language. See [33] for a list of typical stop words.

3.2 Experiment 1 - Word Sense Weighting

Experiments with statistical information retrieval have shown that better performance is achieved by weighting words based on their frequency of use. The most effective weight is usually referred to as TF.IDF, which includes a component based on the frequency of the term in a document (TF) and a component based on the inverse of the frequency within the document collection (IDF) [27]. The intuitive basis for this weighting is that high frequency words are not able to effectively discriminate relevant from non-relevant documents. The IDF component gives a low weight to these words and increases the weight as the words become more selective. The TF component indicates that once a word appears in a document, its frequency within the document is a reflection of the document's relevance.

Words of high frequency also tend to be words with a high number of senses. In fact, the number of senses for a word is approximately the square root of its relative frequency [36].[14] While this correlation may hold in general, it might be violated for particular words in a specific document collection. For example, in the CACM collection the word 'computer' occurs very often, but it cannot be considered very ambiguous.

The intuition about the IDF component can be recast in terms of ambiguity: words which are very ambiguous are not able to effectively discriminate relevant from non-relevant documents. This led to the following hypothesis: weighting words in inverse proportion to their number of senses will give similar retrieval effectiveness to weighting based on inverse collection frequency (IDF). This hypothesis is tested in the first experiment. Using word ambiguity to replace IDF weighting is a relatively crude technique, however, and there are more appropriate ways to include information about word senses in the retrieval model. In particular, the probabilistic retrieval model [33, 10, 15] can be modified to include information about the probabilities of occurrence of word senses. This leads to the second hypothesis tested in this experiment: incorporating information about word senses in a modified probabilistic retrieval model will improve retrieval effectiveness. The methodology and results of these experiments are discussed in the following sections.

3.2.1 Methodology of the weighting experiment

In order to understand the methodology of our experiment, we will first provide a brief description of how retrieval systems are implemented.

Information retrieval systems typically use an inverted file to identify those documents

[14] It should be noted that this is *not* the same as 'Zipf's law', which states that the log of a word's frequency is proportional to its rank. That is, a small number of words account for most of the occurrences of words in a text, and almost all of the other words in the language occur infrequently.

which contain the words mentioned in a query. The inverted file specifies a document identification number for each document in which the word occurs. For each word in the query, the system looks up the document list from the inverted file and enters the document in a hash table; the table is keyed on the document number, and the value is initially 1. If the document was previously entered in the table, the value is simply incremented. The end result is that each entry in the table contains the number of query words that occurred in that document. The table is then sorted to produce a ranked list of documents. Such a ranking is referred to as a 'coordination match' and constitutes a baseline strategy. As we mentioned earlier, performance can be improved by making use of the frequencies of the word within the collection, and in the specific documents in which it occurs. This involves storing these frequencies in the inverted file, and using them in computing the initial and incremental values in the hash table. This computation is based on the probabilistic model, and is described in more detail in the next section.

Our experiment compared four different strategies: coordination match, frequency weighting, sense weighting, and a strategy that combined frequency and sense weighting based on the probabilistic model. Retrieval performance was evaluated using two standard measures: *Recall* and *Precision* [33]. *Recall* is the percentage of relevant documents that are retrieved. *Precision* is the percentage of retrieved documents that are relevant. These measures are presented as tables of values averaged over the set of test queries.

3.2.2 Results of weighting experiment

Table 3 shows a comparison of the following search strategies:

Coordination match: This is our baseline; documents are scored with respect to the number of words in the query that matched the document.

Frequency weighting: This is a standard TF.IDF weighting based on the probabilistic model. Each document is ranked according to its probability of relevance, which in turn is specified by the following function:

$$g(\mathbf{x}) = \sum_{i \in query} tf_i \log \frac{p_i(1-q_i)}{(1-p_i)q_i} \quad (1)$$

where $\mathbf{x}$ is a vector of binary terms used to describe the document, the summation is over all terms in the query, tf_i is the probability that term i is used to index this document, p_i is the probability that term i is assigned to a random document from the class of relevant documents, and q_i is the probability that term i is assigned to a random document from the class of non-relevant documents. These probabilities are typically estimated using the normalized frequency of a word in a document for tf_i, the relative frequency of term i in the collection for q_i, and a constant value for p_i. Using these estimates, ranking function (1) is a sum of TF.IDF weights, where the TF weight is tf_i, and the IDF weight is (approximately) $\log \frac{1}{q_i}$.

Sense weighting: Ranking function (1) is used, but the IDF component is replaced by a sense weight. This weight was calculated as $\log \frac{1}{w_i}$, where w_i is the number of senses of term i in the dictionary normalized by the maximum number of senses for a word in the dictionary; if a word does not appear in the dictionary, it is assumed to have only one sense.

Combined: This is a modification of frequency weighting to incorporate a term's degree of ambiguity. Ranking function (1) assumes that the probability of finding a document representation $\mathbf{x}$ in the set of relevant documents is (assuming independent terms)

$$\prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i}$$

where n is the number of terms in the collection. A similar expression is used for non-relevant documents. Since we are primarily interested in word senses that match query senses, a possible modification of this ranking function would be to compute the probability that the terms in $\mathbf{x}$ represent the correct word sense. For a given term, this probability is $p_i p_{is}$, where p_{is} is the probability of a correct sense. We estimate p_{is} by the inverse of the number of senses for term i, which assumes that each sense is equally likely. The resulting ranking function, which is a minor modification of function (1), is

$$g(\mathbf{x}) = \sum_{i \in query} tf_i \log \frac{p_i(1 - q_i p_{is})}{(1 - p_i p_{is}) q_i} \qquad (2)$$

The table shows the precision at ten standard levels of recall. In the case of the CACM collection, 45 of the original 64 queries were used for this experiment.[15] The results show that the first hypothesis holds in the TIME collection, but not in the CACM collection. The results for sense weighting in the CACM collection are nearly the same as no weighting at

[15] Although the collection contains 64 queries, only 50 are usually used for retrieval experiments. This is because some of the queries do not have any relevant documents, and because some are too specific (they request articles by a particular author). Five additional queries were omitted from our experiment because of an error.

Recall	**CACM** Precision (45 queries) coord	 freq	 sense	 comb.	**TIME** Precision (45 queries) coord	 freq	 sense	 comb.
10	42.7	52.9	40.0	53.0	59.7	63.4	62.0	64.0
20	27.5	37.9	29.9	37.6	57.1	60.3	59.7	61.1
30	21.1	30.9	22.6	31.6	54.9	58.3	57.3	60.7
40	17.4	26.1	16.6	27.1	50.6	55.5	53.6	57.1
50	14.8	22.0	12.9	23.0	49.2	53.5	53.2	54.5
60	11.3	18.5	9.0	18.7	39.1	47.4	46.2	48.3
70	7.7	10.9	5.0	10.3	35.0	44.8	43.1	46.0
80	6.1	7.5	4.0	7.2	33.4	43.7	42.4	44.9
90	4.8	6.3	3.4	6.1	27.9	36.7	35.8	38.3
100	4.5	4.9	2.5	4.8	27.6	36.0	35.4	37.5

Table 3: Weighting Results for the CACM and TIME collections. The Precision is shown for ten standard levels of Recall. The first column (**coord**) is a baseline – no weighting. The next three columns reflect different weighting strategies: one based on term frequency, (**freq**), one based on degree of ambiguity (**sense**), and the last one is a combination of the two (**combined**).

all (the **coord** result), whereas in the TIME collection, sense weighting and IDF weighting give similar results.

The second hypothesis also holds in the TIME collection, but not in the CACM collection. The modified probabilistic model gave small effectiveness improvements for TIME (**comb** vs. **freq**), but in the CACM collection made virtually no difference. This is not unexpected, given the inaccuracy of the assumption of equally likely senses. Better results would be expected if the relative frequencies of senses in the particular domains were known.

3.2.3 Analysis of weighting experiment

The poor performance of sense weighting for the CACM collection raises a number of questions. According to Zipf, the number of senses should be strongly correlated with the square root of the word's frequency. We generated a scatterplot of senses vs. postings[16] to see if this was the case, and the result is shown in Figure 1. The scatterplot shows that most of the query words appear in a relatively small number of documents. This is not surprising; users will tend to use words that are fairly specific. As we expected, it also shows that there are several words that do not have many senses, but which appear in a large number of documents. What *is* surprising is the large number of words that are of high ambiguity and low frequency. We examined those words and found that about a third of them were general vocabulary words that had a domain specific meaning. These are words such as: 'passing' (as in 'message passing'), 'parallel', 'closed', 'loop', 'address', etc. The CACM collection constitutes a sublanguage in which these words generally only occur with a domain-specific sense. We also found several cases where the word was part of a phrase that has a specific meaning, but in which the words are highly ambiguous when considered in isolation, (e.g. 'back end', or 'high level').

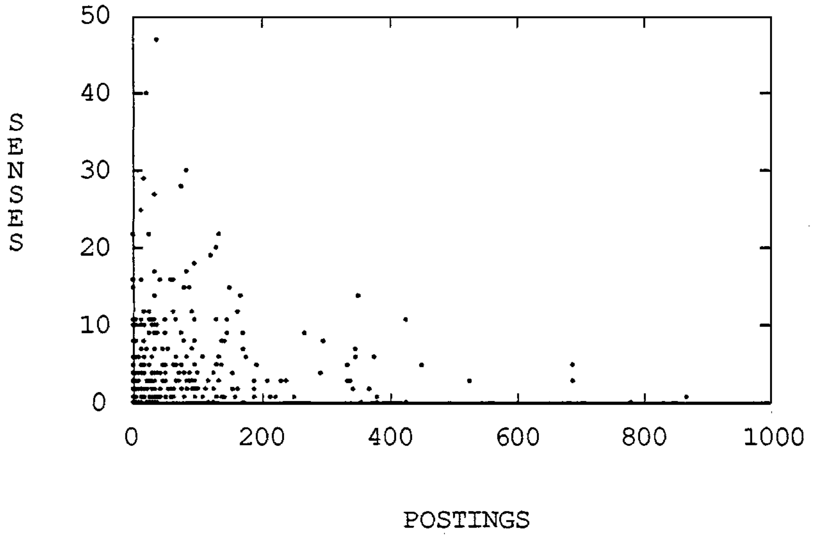

Figure 1: Scatterplot for the CACM queries

These same effects were also noticed in the TIME collection, although to a much smaller

[16] 'postings' refers to the number of documents in which a word appears; we used this value instead of frequency because it is the value used in the calculation of the IDF component. It is a close approximation to the actual word frequency in the CACM collection because the documents are only titles and abstracts.

degree. For example, the word 'lodge' almost always occurs as a reference to 'Henry Cabot Lodge' (although there is an instance of 'Tito's Croation hunting lodge').[17] We found that the TIME collection also had problems with phrases. The same phrase that caused a problem in CACM, 'high level', also appears in TIME. However, when the phrase appears in CACM, it usually refers to a high level programming language; when it appears in TIME, it usually refers to high level negotiations.

Another factor which contributed to the poor results for CACM is the use of common expressions in the CACM queries; these are expressions like:[18] 'I am interested in ...', 'I want articles dealing with ...', and 'I'm not sure how to avoid articles about ...'. While some of these words are eliminated via a stop word list ('I', 'in', 'to'), words such as 'interest', 'sure', and 'avoid' are highly ambiguous and occur fairly infrequently in the collection. None of the queries in the TIME collection included these kind of expressions.

Some of the effects that caused problems with the CACM and TIME collections have also been noticed by other researchers. Keen noticed problems in the ADI collection (a collection of text on documentation) involving homonyms and inadequate phrasal analysis [19]. For example, the word 'abstract' was used in a query in the sense of 'abstract mathematics', but almost always appeared in the collection in the sense of a document summary.[19] The problem with common expressions was also noted by Sparck-Jones and Tait: 'one does not, for example, want to derive a term for 'Give me papers on' ... They [non-contentful parts of queries] are associated with undesirable word senses ...' [32].

3.3 Experiment 2 - Word Sense Matching

Our experiments with sense weighting still left us with the question of whether indexing by word senses will yield a significant improvement in retrieval effectiveness. Our next experiment was designed to see how often sense mismatches occur between a query and a document, and how good a predictor they are of relevance. Our hypothesis was that a mismatch on a word's sense will happen more often in a non-relevant document than in a relevant one. In other words, incorrect word senses should not contribute to our belief that the document is relevant. For example, if a user has a question about 'foreign policy', and the document is about 'an insurance policy', then the document is not likely to be relevant (at least with respect to the word 'policy').

[17] The TIME collection dates from the early 60's.

[18] Note that since full-text systems do not pay any attention to negation, a query that says 'I'm not sure how to avoid articles about ...', will get exactly those articles as part of the response.

[19] The exact opposite problem occurred with the CACM collection; one of the queries referred to 'abstracts of articles', but 'abstract' is often used in the sense of 'abstract data types'.

	CACM	**TIME**
Queries examined	45	45
Words in queries	426	335
Words not in LDOCE	37 (8.7%)	80 (23.9%)
Domain specific sense	45 (10.5%)	6 (1.8%)
Marginal sense	50 (11.7%)	8 (2.4%)

Table 4: Statistics on word senses in test collection queries

To test our hypothesis we manually identified the senses of the words in the queries for both collections. These words were then manually checked against the words they matched in the top ten ranked documents for each query (the ranking was produced using a probabilistic retrieval system). The number of sense mismatches was then computed, and the mismatches in the relevant documents were identified. A subset of 45 of the TIME queries were used for this experiment, together with the 45 CACM queries used in the sense weighting experiment. The TIME queries were chosen at random.

3.3.1 Results of sense matching experiment

Table 4 shows the results of an analysis of the queries in both collections.[20] For the CACM collection, we found that about 9% of the query words do not appear in LDOCE at all, and that another 22% are used either in a domain-specific sense, or in a sense that we considered 'marginal' (i.e., it violated semantic restrictions, or was used in a sense that was somewhat different from the one listed in the dictionary). For example, we considered the following words to be marginal: 'file', 'language', 'pattern', and 'code'; we will discuss such words in more detail in the next section. For the TIME collection the results were quite different. About 24% of the query words were not found in LDOCE, and approximately 4% were used in a domain-specific or marginal sense.

Table 5 shows the result of comparing the query words against the occurrences of those words in the top ten ranked documents. The query words that appeared in those documents are referred to as 'word matches'; they should not be confused with the *senses* of those words. If the *sense* of a query word is the same as the sense of that word in the document, it will be referred to as a 'sense match' (or conversely, a 'sense mismatch').

[20] The numbers given refer to word tokens in the queries. The percentages for word types are similar.

CACM

	All Docs	Relevant Docs
Number	450	116 (25.8%)
Word Matches	1644	459 (27.9%)
Clear Sense Mismatches	116	8 (7.0%)
Technical-General Mismatches	96	6 (6.3%)

TIME

	All Docs	Relevant Docs
Number	450	101 (22.5%)
Word Matches	1964	529 (26.9%)
Clear Sense Mismatches	166	20 (12.1%)
Number of hit+mismatches	127	29 (22.8%)

Table 5: Results of word sense matching experiments. *Word Matches* refers to the occurrences of query words in a document. *Clear Sense Mismatches* are the number of Word Matches in which the sense used in the query does not match the sense used in the document. *Technical-General Mismatches* are the number of Word Matches in which it was difficult to determine whether the senses matched due to the technical nature of the vocabulary; these rarely occurred in the TIME collection. *Hit+Mismatches* are the additional Clear Sense Mismatches that occurred in documents in which there was also a sense *match*; these rarely occurred in the CACM collection due to the length of the documents. The percentages in the Relevant Docs column refer to the number of Relevant Docs divided by All Docs.

The table indicates the number of word matches that were clearly a sense mismatch (e.g., 'great deal of interest'/'dealing with'). Occasionally we encountered a word that was extremely ambiguous, but which was a mismatch on part-of-speech (e.g., 'use'/'user'). It was difficult to determine if these words were being used in distinct senses. Since these words did not occur very often, they were not considered in the assessment of the mismatches.

A significant proportion of the sense mismatches in both collections was due to stemming (e.g., 'arm'/'army', 'passive'/'passing', and 'code'/'E. F. Codd'). In the CACM collection this accounted for 39 of the 116 mismatches, and 28 of the 166 mismatches in the TIME collection.

Each collection also had problems that were specific to the individual collection. In the CACM collection we encountered difficulty because of a general vocabulary word being used with a technical sense (e.g., 'process' and 'distributed'). These are labeled 'technical-general mismatches'. There were 20 sense mismatches that we included in the 'clear mismatch' category despite the fact that one (or both) of the words had a technical sense; this was because they clearly did not match the sense of the word in the query (e.g., 'parallels between problems'/'parallel processing', 'off-line'/'linear operator', 'real number'/'real world'). The technical/general mismatches were cases like 'probability distribution' versus 'distributed system' in which it was difficult for us to determine whether or not the senses matched. Technical-general mismatches rarely caused a problem in the TIME articles. In contrast, the TIME collection sometimes contained words that were used in several senses in the same document, and this rarely occurred in CACM. The number of sense mismatches that occurred in documents in which a sense match *also* occurred are labeled 'hit+mismatches'; 'clear sense mismatches' only includes mismatches in which *all* senses of the word were a mismatch.

For each collection the results are broken down with respect to all of the documents examined, and the proportion of those documents that are relevant.

3.3.2 Analysis of the sense matching experiment

There are a number of similarities and differences between the two test collections. In the queries, about 70% of the words in both collections were found in the dictionary without difficulty. However, there are significant differences in the remaining 30%. The TIME queries had a *much* higher percentage of words that did not appear in the dictionary at all (23.9% versus 8.7%). An analysis showed that approximately 98% of these words were proper nouns (the Longman dictionary does not provide definitions for proper nouns). We compared the words with a list extracted from the Collins dictionary,[21] and found that all of them were in-

[21] This was a list composed of headwords that started with a capital letter.

	CACM	TIME
Connotation:	'parallel' (space vs. time), 'file', 'address', 'window'	'aid' (monetary implication), 'suppress' (political overtones)
Semantic restrictions:	human vs. machine	human vs. country
Too general:		'relationship'
Part-of-speech:	'sort', 'format', 'access'	'shake-up'
Overspecified entry:	'tuning', 'hidden'	
Phrasal lexemes:	'back end', 'context free', 'outer product', 'high level'	'United States', 'left wing', 'hot line', 'high level'

Table 6: Reasons for difficulties in sense match assessment

cluded in the Collins list. We feel that a dictionary such as Longman should be supplemented with as large a list of general usage proper nouns as possible. Such a list can help identify those words that are truly domain specific.

The two collections also showed differences with respect to the words that were in the dictionary, but used in a domain specific sense. In the CACM collection these were words such as 'address', 'closed', and 'parallel' (which also accounted for different results in our previous experiment). In the TIME collection this was typically caused by proper nouns (e.g., 'Lodge' and 'Park' as people's last names, 'China' as a country instead of dinnerware).

There were many instances in which it was difficult to determine whether a word in the document was a mismatch to the word in the query. We considered such instances as 'marginal', and the reasons behind this assessment provide a further illustration of differences as well as similarities between the two collections. These reasons are given in Table 6, and are broken down into 'connotation', 'semantic restrictions', 'too general', 'part-of-speech', 'overspecified entry', and 'phrasal lexeme'. The reasons also account for the entries in Table 4 that were labeled 'marginal sense'; these are query words that were not an exact match for the sense given in the dictionary.

In the CACM collection, differences in connotation were primarily due to a general vocabulary word being used in a technical sense; these are words like 'file', 'address', and 'window'. In the TIME collection the differences were due to overtones of the word, such as the implication of money associated with the word 'aid', or the politics associated with the word 'suppress'. Semantic restriction violations occurred when the definition specified that a verb required a human agent, but a human agent was not used in the given context. This was due to the use of computers as agents in the CACM collection, and the use of countries as agents in the TIME collection. Both TIME and CACM use words with a part-of-speech different from the one given in the dictionary, but they occur much more often in CACM (e.g., 'sort' as a noun, and 'format' and 'access' as verbs; the TIME collection refers to 'shake-up' as a noun although the dictionary only lists it as a verb).

Definitions that were too general or too specific were also a significant problem. For example, the word 'relationship' is defined in LDOCE as a 'connection', but we felt this was too general to describe the relationship between countries. There is also another sense that refers to family relationships, but this caused difficulty due to connotation. Definitions were considered too specific if they referred to a particular object, or if they carried an implication of intentionality that was not justified by the context. The former problem is exemplified by 'tuning', which was defined with regard to an engine but in context referred to a database. The latter problem is illustrated by a word like 'hidden' in the context 'hidden line removal'. Interestingly, problems with generality did not occur with CACM, and problems with overly specified entries did not occur with TIME. Finally, as we previously mentioned, there are a number of words that are best treated as phrasal.

Although both collections show a number of differences, the overall result of the experiment is the same: word senses provide a clear distinction between relevant and non-relevant documents (see Table 5). The null hypothesis is that the meaning of a word is not related to judgments of relevance. If this were so, then sense mismatches would be equally likely to appear in relevant and non-relevant documents. In the top ten ranked documents (as determined by a probabilistic retrieval system), the proportion that are relevant for CACM is 25.8% (116/450), and for TIME the proportion is 22.5% (110/450). The proportion of word matches in relevant documents for the two collections is 27.9% and 26.9% respectively. If word meanings were not related to relevance, we would expect that sense mismatches would appear in the relevant documents in the same proportions as word matches. That is, sense mismatches should appear in relevant documents in the same proportion as the words that matched from the queries. Instead we found that the mismatches constitute only 7% of the word matches for the CACM collection, and 12.1% of the word matches for TIME. We evaluated these results using a chi-square test and found that they were significant in both collections ($p < .001$). We can therefore reject the null hypothesis.

We note that even when there were difficulties in assessing a match, the data shows a clear difference between relevant and non-relevant documents. Sense match difficulties are much more likely to occur in a non-relevant document than in one that is relevant. Most of the difficulties with CACM were due to technical vocabulary, and Table 5 shows the proportion of these matches that appear in relevant documents. The difficulties occurred less often with the TIME collection, only 38 instances in all. However, only 4 of those instances are in documents that are relevant.

Our results have two caveats. The first is related to multiple sense mismatches. When a word in a query occurred in a CACM abstract, it rarely occurred with more than one meaning. In the TIME collection, 6.5% of the word matches were 'hit+mismatch'; these were sense mismatches in which the document also contained a sense of the word that *did* match the query. We found that 22.8% of these mismatches occurred in relevant documents versus an expected 26.9%. This does not constitute enough of a difference to reject the null hypothesis. In other words, as long as the article contained at least one occurrence of the correct sense, it was just as likely to be relevant as a document in which *all* occurrences of the word had the correct sense. If *all* of the occurrences of the word in the article had the wrong sense, then the article was significantly less likely to be relevant. However, even in cases of 'hit+mismatch', it might still be useful to know about the mismatches. This is because retrieval performance can be improved by weighting words by their within-document frequency (see Section 4.2). The basis for this weighting is that once a query word occurs in a document, its frequency within the document is an indicator of the word's importance. If some of the instances of the word are sense mismatches, we might discount them as contributing to that frequency on the grounds that they are not indicators of the same sense. We expect that this would have more of an effect on the retrieval of full-text documents than on collections that consist of just titles and abstracts. In addition, it is often the case in full-text documents that we would like to identify which passages are most relevant; it is possible that the sections of the document that contain mismatches do not contribute a relevant passage and could be discounted in such an assessment.

The second caveat relates to the number of sense mismatches we found. The data indicates that sense mismatches constitute 7 to 13% of the word matches in the CACM collection (depending on whether technical-general sense mismatches are included), and about 8.5%[22] of the word matches in the TIME collection. If these results are a reflection of the entire ranking, distinguishing word meanings would probably not make a significant improvement in performance.

One explanation for the high degree of matching is that some senses may occur very frequently; if the sense of the word given in the query is a frequent sense, it would be expected to match the corpus a high percentage of the time. This frequency might be a reflection of the word's distribution in English, or it might be due to the sublanguage of the collection. For example, although the word 'prime' is ambiguous, it almost always occurs in the sense 'prime number' in CACM, and almost always in the sense 'prime minister' in TIME. Another explanation is that the high degree of matching was due to word collocation. The documents we examined were the top ten ranked documents for each query. These documents have the most words in common with the query; because the words are related to each other (by virtue

[22] They actually constitute 15% of the word matches, but 6.5% are 'hit+mismatches' that do not show a difference between relevant and non-relevant documents.

of being part of a query that has an overall meaning), this tends to provide constraints on their meanings. Our hypothesis is that we will be more likely to get a mismatch on documents that have only one word in common with a query than on those in which many words are in common. Approximately 75% of the documents retrieved for the CACM queries have only one word in common with the query. For the TIME collection, about 54% of the documents have only one word in common.[23] We therefore have the potential for eliminating a large number of non-relevant documents.

We tried to conduct an experiment to test whether the above hypothesis was correct, but this proved difficult to do. We wanted to determine whether sense mismatches occur more often in the documents below the top ten, and whether they still give a good separation between those that are relevant and those that are non-relevant. However, the number of words in a query varies considerably, as does the number of documents retrieved in response. We were unable to find a range of ranks which consistently contained relevant documents as well as documents with only a few words in common with the query.

The top ranked documents have a large number of words in common with the query and disambiguation is not likely to have much of an effect on how these documents are ranked. However, we wanted to gain a better understanding of how the separation achieved by word senses would be reflected in improved performance. Because sense mismatches are much more likely to appear in non-relevant documents, we eliminated every document from the top ten that contained a mismatch on any query word. The precision was determined for each set of ten documents and averaged over all of the queries examined. This was compared with the average precision after removing the documents that contain a mismatch. The result was that the precision increased from 26 to 34.5% for the CACM collection, and from 22.4% to 23.3% for the TIME collection.

3.4 Experiment 3 - Word Sense Distribution

What seems necessary is a way to identify those words that are worth disambiguating. Some words are theoretically very ambiguous, but because one of their senses occurs very often, in practice they can be considered as being 'relatively unambiguous'. In addition, we were still left with the question of why sense matches occur so often; was it due to collocation, or to the distribution of senses in the collection? To make this assessment we examined a KWIC index of the entire CACM collection. A KWIC index is a listing of the words in the collection, sorted alphabetically, with each word appearing on its own line along with the context in which it appears. For each word in the queries for the CACM collection, we

[23] This figure is probably lower because of the larger number of documents in the CACM collection, half of which consist of only a title.

examined the KWIC index and made an estimate of the distribution of the senses for that word.[24]

The analysis of the corpus distribution was not easy to do. We needed to consider not only the occurrence of the word form, but also any morphological variations and whether those variations had a sense that was significantly different from the root. Dictionaries recognize these differences and will list a variant separately if it has a distinct meaning. Examples are words like 'essential' and 'essentially', 'multiple' and 'multiplication', and 'product' and 'production'. As with the previous experiment, we needed to be concerned with sense distinctions even if they were not reflected in the dictionary (due to the technical nature of the text). For example, the word 'complex' is conflated with the word 'complexity' by the stemmer used in the retrieval system, and would also be conflated by most morphological analysis routines. However, 'complex' is almost always used in CACM to refer to a complex number, and 'complexity' refers to a concept in theoretical computer science. Such variation was also a problem in the previous experiment, but in this experiment we were faced with every variation of the word throughout the corpus.

3.4.1 Results of the sense distribution experiment

The results of our experiment are given in Table 7 and 8. The first table shows the proportion in which each sense appears in the corpus. For purposes of comparison we also give the distribution of senses within the queries themselves. The data shows that the distribution in the queries is a reflection of the distribution over the entire corpus. Approximately 74% of the senses are either sense 1, sense 2, or domain specific. As we mentioned in the previous section, some of the senses were difficult to assess and they were categorized as 'marginal'. The domain specific senses are either words that are not in LDOCE at all (e.g., 'stochastic', 'database', 'robotics'), which rarely have more than one sense, or else are words like 'address' and 'loop', which are in LDOCE but are being used in a technical sense. The most noticable difference between the queries and the corpus is that the corpus contains a higher proportion of general vocabulary words that are used in a technical sense (and a correspondingly lower proportion of words that are not found in the dictionary at all).

The data in Table 7 also provides empirical evidence that the senses in the Longman dictionary are ordered by frequency. The first sense listed in the dictionary constitutes approximately 40% of all senses in the queries and in the corpus. This might have been the result of many words with only one sense, but Table 8 indicates that this was the case for only 14% of the query words. Some dictionaries do *not* order their senses by frequency, but

[24] We did not analyze the sense distribution for the TIME collection.

	Queries (n=306)	**Corpus** (n=480)
Sense 1	132 (43.1%)	198 (41.3%)
Sense 2	31 (10.1%)	60 (12.5%)
Sense 3	17 (5.6%)	35 (7.3%)
Sense 4	6 (2.0%)	13 (2.7%)
Sense 5	5 (1.5%)	8 (1.7%)
Sense > 5	14 (4.6%)	10 (2.1%)
Marginal	38 (12.4%)	45 (9.4%)
Domain Specific (in LDOCE)	27 (8.8%)	67 (14%)
Domain Specific (non LDOCE)	36 (11.8%)	34 (7.1%)

Table 7: Distribution of senses within CACM queries and corpus. Each row indicates the number of occurrences of the first sense listed in the dictionary, the second sense, etc. (n is the number of unique senses in each dataset).

	Corpus (n=295)	**LDOCE** (n=295)
One sense	133 (45.1%)	43 (14.6%)
Two senses	83 (28.1%)	45 (15.3%)
Three senses	26 (8.8%)	32 (10.8%)
Four senses	12 (4.1%)	26 (8.8%)
Five senses	2 (0.7%)	18 (6.1%)
Six senses	2 (0.7%)	10 (3.4%)
More than six senses	0	84 (28.5%)
Not in LDOCE	37 (12.5%)	37 (12.5%)

Table 8: Number of senses for CACM query words. Each row indicates the number of occurrences of words with the given number of senses (n is the number of unique words in the queries; morphological variants are conflated unless the variant appears in the dictionary).

by the chronological order in which they entered the language. Although it is not shown here, the TIME queries show a very similar breakdown in the proportion of senses. These proportions can be useful in a disambiguation system if we are otherwise unable to determine which sense is correct.

Our examination of the KWIC index also provided us with data about how many senses were observed for each word, and the relative proportion for each sense. In addition, we determined the number of senses each word has in the dictionary. These figures are given in Table 8 and provide a comparison of the number of senses indicated by the dictionary versus the number that were actually observed. Some words had more senses than are indicated by the dictionary due to their use in technical senses. The observed senses are a lower bound on the number that are actually in the corpus; some distinctions may have been glossed over because of the large number of senses we examined. We also ignored any senses that appeared less than 1 percent of the time. Our aim was to obtain a rough indication of how senses were distributed in the corpus, and to determine whether the results of our previous experiment were due to collocation or to sense distribution.

Table 8 shows that there is considerable ambiguity even in a specialized database; over 40% of the query words were found to have more than one sense. Although words in the corpus appear to have a mean of 4.4 senses based on a dictionary look-up (see Table 2), the mean number of senses based on our observations is only 1.6. However, the first mean includes any uses that are idiomatic, and these are fairly rare in practice. We noted a few idiomatic uses in our examination of the KWIC index, but they made up such a small percentage of the overall uses that they were not counted. The dictionary also includes separate senses for each phrasal verb[25] as well as any distinctions within that category. These uses can account for a large number of the senses attributed to a word. We were not able to determine how often phrasal verbs occurred because they are mixed in with non-phrasal uses of the verb, and the overall frequency is very high.

Finally, we identified the proportion of words that would be worth disambiguating. These are words that fall into one of two categories:

1. The word does not have any senses in a skewed distribution (where we defined 'skewed' to mean that one of the senses occurs 80% of the time or more). We term these words 'uniformly ambiguous'.

2. The word has a skewed distribution, but the query sense is one of the senses in the minority.

[25] A phrasal verb is a verb that is followed by a preposition or an adverb (jointly referred to as 'particles'), and in which the two words together make up one lexical unit (e.g., 'take up', 'look at', 'give in', etc.). The particle may or may not be adjacent to the verb, and occasionally may even be omitted.

Our analysis showed that 46 words (15.6%) were in the first category, and 26 words (8.8%) were in the second category. However, these percentages refer to the breakdown of word *types*, and our previous experiment was concerned with word *tokens*. We examined the number of tokens for these words and found that the proportion of tokens they represented was almost identical.

We repeated our previous experiment in which we eliminated every document from the top ten that contained a mismatch on any query word. However, this time we only considered the subset of the mismatches that involved the words we identified as 'worth disambiguating'. The result was that the average precision increased from 26% to 28.6%. This was not as much of an improvement as in the previous experiment because we only considered a subset of the mismatches, and because we did not consider all of the mismatches caused by stemming.

3.4.2 Analysis of sense distribution experiment

The aim of this experiment was to determine the reason behind the high degree of matching in our previous results. Was it due to the effect of collocation, or to the distribution of senses in the corpus? The data shows that both effects were at work. Approximately 24% of the word occurrences have a likelihood of a mismatch (either they are uniformly ambiguous, or they are relatively unambiguous but have a query sense with a minority usage). Instead we found that sense mismatches constitute only 13% of the pairings between a query word and a word in a document.[26]

As we mentioned earlier, approximately 75% of the documents retrieved for the CACM queries have only one word in common with the query. We can expect that distinguishing the meanings of these words would remove many of them from the ranking and therefore lead to an improvement in performance. This improvement would be most noticeable at the lowest levels of recall. Improvement at higher levels of recall might also be possible if we augment the query using a thesaurus, but only include words that are used with a relevant sense.

4 Conclusion

Previous work on lexical ambiguity has dealt with only a small number of words, a restricted range of senses, or both. Although the information retrieval literature has noted that word sense ambiguity is a problem, very little work has been done to determine how often it occurs and how much impact it has on performance.

[26] This figure includes mismatches due to words in the general vocabulary being used with a technical sense.

Our first experiment was concerned with weighting words by the number of senses they have. This was done in order to gain a better understanding of the relationship between word frequency and ambiguity. The experiment showed that word sense weighting improved retrieval effectiveness by a small amount in one collection, and made no difference in the other. We determined that this was partially due to general vocabulary words being used in a technical sense, and this led to the observation that an anomalous frequency distribution can be useful for detecting domain specific word senses.

Our next experiment was concerned with determining how often the sense of words in a query match the senses of those words in a document. This experiment shows that there is a very strong correlation between the meaning of words in a query, the meaning of those words in a document, and judgments of relevance. Word sense mismatches are far more likely to appear in non-relevant documents than in those that are relevant. Word sense matches were, however, very frequent, and the reason for the high degree of matching was not clear. It may be due to the effect of word collocation in the set of documents we examined (which had the most words in common with the query), or the distribution of senses in the corpus. We analyzed a KWIC index of the corpus and determined that both factors were contributing. Approximately 24% of the words were likely to have a mismatch, but mismatches make up only 13% of the query-word/document-word pairs.

Word sense mismatches also show a significant difference between the two collections we examined. In one collection, if a word appears in a document at all, it almost always appears with the same sense. In the other collection there were a number of documents in which a word occurs with several senses. We found that as long as the document contains at least one occurrence of the sense of a word from the query, the likelihood of relevance is not affected. If *all* of the senses of those words are a mismatch, the document is unlikely to be relevant.

We believe that resolving word senses will have the greatest impact on a search that requires a high level of recall. This is because such searches retrieve many documents that have only one word in common with the query. Lexical ambiguity is not a significant problem in documents that have a large number of words in common with a query. Nevertheless, there are several reasons why we believe that disambiguation is worthwhile. First, the test collections we used are both on particular subject areas; we expect that with other text databases, such as patents or dissertation abstracts, ambiguity will be more of a problem. Second, the words in the queries were matched against the words in the text via a process called "stemming" (essentially truncation of the word endings). This process does not capture all of the variants a word can have, and thus some documents will not be retrieved due to a failure to match on a variant (for example, 'actor' will not match 'act' or 'actress'). Such variants are based not on the word, but on the *sense* of the word. Third, a query often does not contain all of the words that might be used to find relevant documents. Disambiguation has the potential for improving precision for low recall searches via the use of a sense-disambiguated thesaurus. Fourth, distinguishing word senses may be useful for highlighting the relevant passages in full-text documents. Finally, the senses of the words is only one factor affecting relevance. The *relationships* that those words have to one another is also important. Determining these relationships is likely to require the use of a natural language parser, and knowing the senses in which the words are used serves as an important constraint on the parse. Although the words may only be used with a small number of senses (relative to the number they have in the dictionary), we do not know in advance which *particular* senses will be used within a given collection of text. Word sense disambiguation is also important in other areas of natural language processing such as machine translation and text critiquing.

5 Current and Future Work

The work reported in this paper was done in order to get a better understanding of lexical ambiguity, and the effect that it has on information retrieval. An accurate assessment of the impact of word senses on performance will require the implementation of a system for disambiguation. We also wish to determine which aspects of a word's meaning have the greatest benefit in determining its sense. Our approach is based on treating the information associated with dictionary senses (part-of-speech, subcategorization, word collocations, etc.) as multiple sources of evidence. We will be investigating how well each source discriminates senses, how well it can be identified with a word in context, and how much improvement it makes in the performance of a retrieval system. The sources will first be examined independently, and they will then be combined to see how much improvement is gained through consensus.

Acknowledgments

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 90-0110, and by NSF Grant IRI-8814790. We wish to thank Longman Group, Ltd. for making the machine-readable version of the Longman dictionary available to us, and to Collins Publishers and Robert France for providing us with the list of proper nouns from the Collins English Dictionary. We would also like to thank David Lewis and Howard Turtle for making comments on a draft version of this paper.

References

[1] Amsler R., "The Structure of the Merriam Webster Pocket Dictionary", PhD Dissertation, University of Texas at Austin, 1980.

[2] Anthony E, "An Exploratory Inquiry into Lexical Clusters", *American Speech*, Vol 29(3), pp. 175–180, 1954.

[3] Atkins B., "Semantic ID Tags: Corpus Evidence for Dictionary Senses", Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary, pp. 17-36, 1987.

[4] Becker B., "Sense Disambiguation using the *New Oxford English Dictionary*", Masters Thesis, University of Waterloo, 1989.

[5] Belkin N. and Croft W. B., 'Retrieval Techniques', in *Annual Review of Information Science and Technology (ARIST)*, Vol. 22, pp. 109-145, 1987

[6] Black E., "An Experiment in Computational Discrimination of English Word Senses", *IBM Journal of Research and Development*, Vol. 32, No. 2, pp. 185–194, 1988.

[7] Chodorow M., Ravin Y., and Sachar H., "Tool for Investigating the Synonymy Relation in a Sense Disambiguated Thesaurus", *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pp. 144–151, 1988.

[8] Choueka Y. and Lusignan S., "Disambiguation by Short Contexts", *Computers and the Humanities*, Vol. 19, pp. 147–157, 1985.

[9] Cottrell G. and Small S., "A Connectionist Scheme for Modeling Word Sense Disambiguation", *Cognition and Brain Theory*, Vol. 6, No. 1, pp. 89–120, 1983.

[10] Croft W.B., "Experiments with Representation in a Document Retrieval System", *Information Technology: Research and Development*, Vol. 2, pp. 1–21, 1983.

[11] Cullingford R. and Pazzani M., "Word-Meaning Selection in Multiprocess Language Understanding Programs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 4, 1984.

[12] Dahlgren K., *Naive Semantics for Natural Language Understanding*, Kluwer Academic Publishers, 1988.

[13] Earl L., "Use of Word Government in Resolving Syntactic and Semantic Ambiguities", *Information Storage and Retrieval*, Vol. 9, pp. 639–664, 1973.

[14] Fox E., Nunn G., and Lee W., "Coefficients of Combining Concept Classes in a Collection", *Proceedings of the Eleventh International Conference on Research and Development in Information Retrieval*, pp. 291–308, 1988.

[15] Fuhr N., "Models for Retrieval with Probabilistic Indexing", *Information Processing and Management*, Vol. 25, No. 1, pp. 55–72, 1989.

[16] Hayes P., "Some Association-based Techniques for Lexical Disambiguation by Machine", Ph.D. Dissertation, published as Technical Report No. 25, Dept. of Computer Science, University of Rochester, 1977.

[17] Helm S., "Closer Than You Think", *Medicine and Computer*, Vol. 1, No. 1., 1983

[18] Hirst G., "Resolving Lexical Ambiguity Computationally with Spreading Activation and Polaroid Words", in *Lexical Ambiguity Resolution*, Small, Cottrell and Tannenhaus (eds), Morgan Kaufmann Press, 1988.

[19] Keen E., "An Analysis of the Documentation Requests", in *The SMART Retrieval System*, G. Salton (ed), Prentice-Hall, 1971.

[20] Kelly E. and Stone P., *Computer Recognition of English Word Senses*, North-Holland Publishing, 1975.

[21] Krovetz R., "Lexical Acquisition and Information Retrieval", in *Lexical Acquisition: Building the Lexicon using On-Line Resources*, U. Zernik (ed), LEA Press, forthcoming.

[22] Krovetz R. and Croft W.B., "Word Sense Disambiguation Using Machine Readable Dictionaries", *Proceedings of the Twelfth International Conference on Research and Development in Information Retrieval*, pp. 127–136, 1989.

[23] Lesk M., "Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a Pine Cone from an Ice Cream Cone", *Proceedings of SIGDOC*, pp. 24–26, 1986.

[24] Masterman M., Needham R.M., Sparck-Jones K., and Mayoh B., "Agricola Incurvo Terram Dimovit Aratro", Report ML84, Cambridge Language Research Unit, 1957, Reprinted 1986.

[25] Proctor P., *Longman Dictionary of Contemporary English*, Longman, 1978.

[26] Salton G., *Automatic Information Organization and Retrieval*, McGraw-Hill, 1968.

[27] Salton G. and McGill M., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[28] Sedlow S. and Mooney D., "Knowledge Retrieval from Expert Systems: II. Research Results", *Proceedings of the 51st Annual Meeting of the American Society of Information Science*, pp. 209–212, 1988.

[29] Slator B., "Lexical Semantics and Preference Semantics Analysis", Ph.D. Dissertation, Report MCCs-88-143, New Mexico State University, 1988.

[30] Small S. and Rieger C., "Parsing and Comprehending with Word Experts (a Theory and its Realization)", in *Strategies for Natural Language Processing*, Lehnert and Ringle (eds), LEA Press, pp. 89–147, 1982.

[31] Small S., Cottrell G., and Tannenhaus M. (eds), *Lexical Ambiguity Resolution*, Morgan Kaufmann, 1988.

[32] Sparck Jones K. and Tait J., "Automatic Search Term Variant Generation", Journal of Documentation, Vol. 40, No. 1, 1984, pp. 50–66.

[33] Van Rijsbergan C. J., *Information Retrieval*, Butterworths, 1979.

[34] Weiss S., "Learning to Disambiguate", *Information Storage and Retrieval*, Vol. 9, pp. 33–41, 1973.

[35] Wilks Y., Fass D., Guo C-M., McDonald J., Plate T., and Slator B., "A Tractable Machine Dictionary as a Resource for Computational Semantics", in *Computational Lexicography for Natural Language Processing*, Boguraev and Brisoce (eds), Longman, 1989.

[36] Zipf G., "The Meaning-Frequency Relationship of Words", *Journal of General Psychology*, Vol. 33, pp. 251–266, 1945.