Retrieving with good sense

Mark Sanderson
Department of Information Studies
The University of Sheffield
Western Bank, Sheffield S10 2TN, UK

Email: m.sanderson@sheffield.ac.uk
Tel: +44 (0) 114 22 22648
Fax: +44 (0) 114 27 80300

1 Abstract

Although always present in text, word sense ambiguity only recently became regarded as a problem to information retrieval which was potentially solvable. The growth of interest in word senses resulted from new directions taken in disambiguation research. This paper first outlines this research and surveys the resulting efforts in information retrieval. Although the majority of attempts to improve retrieval effectiveness were unsuccessful, much was learnt from the research. Most notably a notion of under what circumstance disambiguation may prove of use to retrieval.

2 Introduction

Managers of an online news retrieval system found that queries about a particular person were causing problems with their Information Retrieval (IR) system[1]. Users had tried to retrieve articles about a certain British Prime Minister using the query "major". Many articles about "Prime Minister John Major MP" were retrieved, but in addition many more articles were retrieved where "major" was used as an adjective or as the name of a military rank. Although people are able to interpret the correct sense of a word without conscious effort (except perhaps in the context of jokes), they appear equally able at forgetting that their query words can be ambiguous.

Automatic word sense disambiguation has long been studied: Gale, Church and Yarowsky *[Gale 92]* cite work dating back to 1950. For many years, disambiguators could only accurately disambiguate text in limited domains or over a small vocabulary. In recent years, however, the situation changed with large improvements in scalability resulting in the possibility of applying a disambiguator to accurately resolve the senses of words in a large heterogeneous corpus. Given this ability, researchers began to imagine using disambiguators to process whole document collections. With a more accurate representation and a query also marked up with word senses (either automatically identified or marked up by the user), researchers believed that the accuracy of retrieval would have to improve. This paper presents a survey of the work on word sense disambiguation and information retrieval

First, an overview of disambiguation research is presented, the aim of which is to illustrate the basic principles behind disambiguation algorithms along with a discussion of the issues surrounding their testing. Next, the attempts at explicitly using disambiguation in IR are presented, followed by a discussion citing a number of analysis papers on the subject. The paper concludes with an examination of other approaches to improving retrieval effectiveness that use knowledge of senses, but in a less explicit manner[2].

3 Approaches to disambiguation

The overview of disambiguation research presented here is divided into two classes: disambiguation based on manually created rules; and disambiguation exploiting evidence derived from large corpora.

[1] This was a personal communication with the author.

[2] Note that the research in this review is not presented in strict chronological order. The first mention of a particular innovation may not correspond to its first description.

3.1 Disambiguation based on manually generated rules

An early example of disambiguation is the research of Weiss *[Weiss 73]* who manually constructed a set of rules to disambiguate five words. These rules were of two types, general context rules, and template rules. A general context rule would state that an ambiguous word occurrence had a certain sense if a particular word appeared near that ambiguous word. For example, if the word "type" appeared near to "print" it most likely meant a small block of metal bearing a raised character on one end. Template rules stated that a word occurrence was a certain sense if a particular word appeared in a specific location relative to that occurrence (such an occurrence is often called a local collocation). For example, if "of" appeared immediately after the word "type", the sense of that occurrence was likely to mean a subdivision of a particular kind of thing.

Following limited testing, Weiss found that template rules were better at determining sense than the context rules, and so applied them first. To create these rules, Weiss examined 20 occurrences of an ambiguous word, and then tested these manually created rules on a further 30 occurrences. These tests were performed for five ambiguous words. The accuracy of the resulting disambiguator was of the order of 90%. Weiss examined the erroneous disambiguations and found them to be mostly idiomatic uses.

A larger disambiguator was built by Kelly & Stone *[Kelly 75]* who manually created a set of rules for 6,000 words. They consisted of contextual rules similar to those created by Weiss. In addition, there were rules for checking certain grammatical aspects of a word occurrence: the grammatical category of a word can be a strong indicator of its sense, for example "the train" and "to train". The grammar and context rules were grouped into sets so that only certain rules were applied in certain situations. Conditional statements controlled the application of rule sets. Unlike Weiss's system, this disambiguator was designed to process a whole sentence at the same time. It could vary the order in which the words of a sentence were disambiguated by stopping the disambiguation of one word, trying to process others, and then returning to the original word to discover if disambiguation could now be completed. The system, however, was not a success and Kelly & Stone reported:

> "...we applied these techniques very energetically to real human language, and it became absolutely clear that such a strategy cannot succeed on a broad scale."

Another approach to disambiguation was tried by Small & Rieger *[Small 82]* using what they called "word experts", which were essentially programs. Their idea was to build an expert for each ambiguous word. When disambiguating words in a sentence the expert of each of the words was invoked. It examined its word's context, make decisions about the possible senses of that word and publicise these decisions to the other experts. If, when processing its evidence, an expert could do no more, it became 'dormant' and waited for other experts in the sentence to publicise their decisions. This additional evidence hopefully provided further clues to the dormant expert to enable it to 'awake' and finish disambiguating its word. There is no mention of testing this disambiguator although Small & Rieger state that they are

> "…convinced that distributed word experts will prove to be the only acceptable course in modelling human language."

Along with this optimism, however, they also make clear the magnitude of the task:

> "the expert for the word 'throw' is currently six pages long … this is large, but it should be ten times that size."

Although Kelly & Stone and Small & Rieger came to different conclusions on the potential success of large scale manually-based disambiguation techniques, there is a common thread in their two works: the effort involved in building this type of disambiguator is significant (possibly insurmountable). Since the late 1980s, disambiguation research placed less emphasis on manually created rules and concentrated instead on automatically generated rules based on sense evidence derived from a machine readable corpus and it is this form of disambiguation that is now discussed.

3.2 Disambiguation using evidence from existing corpora

An early example of corpus based disambiguation is the work of Lesk *[Lesk 88]*. He used the textual definitions of a dictionary to provide evidence for his disambiguator, the workings of which can be illustrated with an example. Suppose we wished to resolve the sense of the occurrence of "ash" in the following sentence.

> There was ash from the coal fire.

To disambiguate "ash", its dictionary definition, taken from Longman's Dictionary of Contemporary English (LDOCE) *[Longman 88]*, was looked up and the individual senses of this word (two in this case) were identified.

- ash (1): The soft grey powder that remains after something has been burnt.
- ash (2): A forest tree common in Britain.

Next the definitions of each of the context words in the sentence (apart from stop words) were looked up.

- coal (1): A black mineral which is dug from the earth, which can be burnt to give heat.
- fire (1): The condition of burning; flames, light and great heat.
- fire (2): The act of firing weapons or artillery at an enemy.

What followed was a process similar to ranked retrieval: the individual dictionary sense definitions of "ash" were regarded as a small collection of documents (a collection of two in this case); and the definitions of the context words, "coal" and "fire", were together regarded as a query. The two sense definitions were ranked by a scoring function based on the number of words co-occurring between a sense's definition and the definitions of all context words. The top ranked definition was chosen to be the sense of this occurrence of "ash". Given the short length of the definition texts, it is questionable how often the word overlap necessary for disambiguation occurred. Lesk performed, what he called, "very brief experimentation" and reported a disambiguation accuracy of between 50% and 70%. He did not make clear how often disambiguation failed due to lack of evidence, although he did acknowledge that definition length was likely to be an important factor in deciding which dictionary to use with the disambiguator. After Lesk, much varied research was conducted using online corpora, a selection of which is now reviewed. (A larger survey can be found in *[Sanderson 97]*.)

3.2.1 Further use of dictionaries

In the example showing Lesk's dictionary based technique, only one word co-occurred between the definition of the context words and the dictionary sense definitions of "ash". Wilks et al *[Wilks 90]* addressed this word overlap problem by using a technique of expanding a dictionary definition with words that commonly co-occurred with the text of that definition. This co-occurrence information was derived from all definition texts in the dictionary. Wilks et al worked with LDOCE, which was intended for people that spoke English as a second language: all its definitions were written using a simplified vocabulary of around 2,000 words. Wilks et al stated that this vocabulary had few synonyms, which would have been a distracting element in the co-occurrence calculations. In an example taken from the paper, a set of co-occurring words for the definitions of the word "bank" was illustrated. For the economic sense, words like "money", "check" and "rob" (amongst others) were found to co-occur; for the geographic sense, "river", "flood" and "bridge" were typical of words found.

Like Lesk's method, the disambiguation process Wilks et al used was similar to ranked retrieval. They tested the accuracy of their disambiguator on the word "bank" as it appeared in around 200 sentences. The disambiguator was judged correct on a sentence if it selected the same sense as the one manually chosen. The senses of "bank" are defined in LDOCE at two levels of lexical granularity. At the fine-grained level LDOCE defines 13 senses. Wilks et al reported that their system selected the correct sense 53% of the time. These senses are grouped into five homographs[3], and at this 'coarser' level, the system correctly selected the homograph of "bank" 85% of the time.

Wilks et al noted that it would be desirable to disambiguate a whole sentence simultaneously. But they pointed out that to exhaustively check every permutation of word sense assignments in a typical sentence would involve examining hundreds of thousands of sense combinations. As a solution, they suggested using the technique of simulated annealing which had been applied successfully to problems that were prone to combinatorial explosion.

3.2.2 Disambiguating simultaneously

Cowie et al. *[Cowie 92]* took up this suggestion. Using similar sense disambiguation techniques to Wilks, they built a disambiguator that attempted to simultaneously resolve all the ambiguous words in a sentence. They tested their disambiguator on a total of 67 sentences and reported an accuracy of 47% when resolving to the LDOCE senses.

[3] Homograph: a word which is spelled the same as another word and might be pronounced the same or differently but which has a different meaning.

Disambiguation based on homographs was performed with an accuracy of 72%. Unfortunately they did not compare their system with Wilks et al's disambiguation results. Thus it is hard to decide on the relative merits of their methods.

An issue that these disambiguation researchers failed to address was the establishment of a reasonable baseline against which accuracy results could be compared. One possible baseline was measuring disambiguation accuracy by just randomly selecting senses. It turns out, however, that there was something better as is illustrated in the testing of the following method.

3.2.3 Manually tagging a corpus

A technique that has proved successful in the related field of *part of speech* tagging *[Garside 87]* is to manually mark up a large text corpus with part of speech tags and then train a statistical classifier to associate features with occurrences of the tags. One of the biggest efforts using this approach in sense disambiguation is the work of Ng & Lee *[Ng 96]* who manually disambiguated 192,000 occurrences of 191 words. When training their classifier on the sense tagged occurrences, it examined the following features:

- the part of speech and morphological form of the sense tagged word;
- the unordered set of its surrounding words;
- local collocations (i.e. words or phrases that always occur in a certain position) relative to it;
- and if the sense tagged word was a noun, the presence of a verb was noted also.

In their tests, Ng & Lee found that the local collocation and part of speech/morphological information were the best features for disambiguation. The surrounding words were less good and the presence of a verb proved to be the least useful. Ng & Lee separated their corpus into training and test sets on an 89%-11% split. Their disambiguator tagged word senses with an accuracy of 63.7%. Ng & Lee also established a baseline accuracy measure using the simplistic strategy of always selecting the most common sense of a word. Using this strategy on their corpus, Ng & Lee measured a baseline of 58.1%, the disambiguator provided only a small improvement. The sense definitions used were from the public domain thesaurus WordNet *[Miller 90]* (see also *[WordNet]*), which on average has 7.8 senses per word for the nouns tested and 12.0 for the verbs: a similar 'granularity' of sense to the definitions used by Wilks and by Cowie. As tempting as it might be, meaningful comparisons between the WordNet and LDOCE based work cannot be made due to the differences in the words being tested, the different corpora those words appeared in and the differences in the sense definitions used.

Perhaps the only common theme to be drawn from the work presented so far is that the accuracy of the disambiguators is not high. The only way in which better accuracy was achieved was to resolve words to the homograph level of sense distinction as shown by Wilks. Yarowsky pursued this strategy of using coarser sense distinctions.

3.2.4 Using Thesauri

A number of researchers have used a thesaurus in disambiguation research. One such example is the work of Yarowsky *[Yarowsky 92]* who used Roget's thesaurus *[Kirkpatrick 88]* along with the Grolier Multimedia Encyclopædia *[Grolier]*. His disambiguator was based on the 1,042 semantic categories into which all words in Roget are placed. These are broad categories covering areas like, tools/machinery or animals/insects. Yarowsky's disambiguator attempted to resolve an ambiguous word to one of these categories. For example, deciding if an occurrence of the word "crane" was the tools/machinery or the animal/insect category.

In acquiring evidence to decide which semantic category (sense) an ambiguous word occurrence should be assigned, a set of clue words, one set for each category, was derived from a part of speech tagged Grolier Encyclopædia. To derive one of these clue word sets for a category, every occurrence of every word in that category was looked up in Grolier and the context of each occurrence (the 100 words surrounding that occurrence) was gathered. For example, the category tools/machinery contains 348 words, which together occurred 30,924 times within Grolier. The contexts of each of these occurrences were gathered and it was from the contexts that the clue words were derived using a term selection process similar to relevance feedback. Yarowsky reported deriving around 3,000 clue words per category. The following is a list of some of the words selected for the category animal/insect:

species, family, bird, fish, breed, cm, animal, tail, egg, wild, common, coat, female, inhabit, eat, nest.

Note as Yarowsky stated in his paper:

> "...these are not a list of members of the category; they are the words which are likely to co-occur with the members of the category."

Disambiguation of a word occurrence involved a comparison between the words in its context and the clue word sets of its possible senses. Sets were ranked based on the term overlap between their clue words and the context. In testing, Yarowsky trained his disambiguator for 12 ambiguous words. Several hundred occurrences of each of these words were manually disambiguated. The accuracy of the disambiguator varied, but on average it resolved word senses with an accuracy of 92%. The test words were selected because they had been used in other disambiguation research and, therefore, some comparison was attempted between this and previous work. However, comparisons were suspect because none of the other researchers had tried to disambiguate using the Roget definitions of word sense. Having only 1,042 senses for the whole English language resulted in generally coarser senses than those used by the others. Although the disambiguator is more accurate than previous research, its task appears to be easier.

3.2.5 Testing disambiguators

As can be seen, one problematic area of disambiguation is the issue of measuring the accuracy of a disambiguator. Unlike IR, there are few 'pre-disambiguated' test corpora publicly available and researchers are often faced with the time consuming task of manually disambiguating all the occurrences of the words they wish to test. It is often not possible to share disambiguated texts across research projects because the definitions of word sense each project uses can differ. For example, WordNet defines 15 senses of "bank" where as LDOCE defines 13. As was seen in the review of research presented above, comparing the quality of disambiguation techniques is hard because it is likely that each technique being examined is measuring its accuracy from different sense definitions. The availability of corpora is improving; the corpus generated from Ng & Lee's work is available through the Linguistic Data Consortium (catalogue number LDC97T12). A sense tagged version of the Brown corpus, called SEMCOR, is available (see *[WordNet]*). There is also a TREC-like effort underway, called SENSEVAL, to compare the accuracy of disambiguators on a common test corpus.

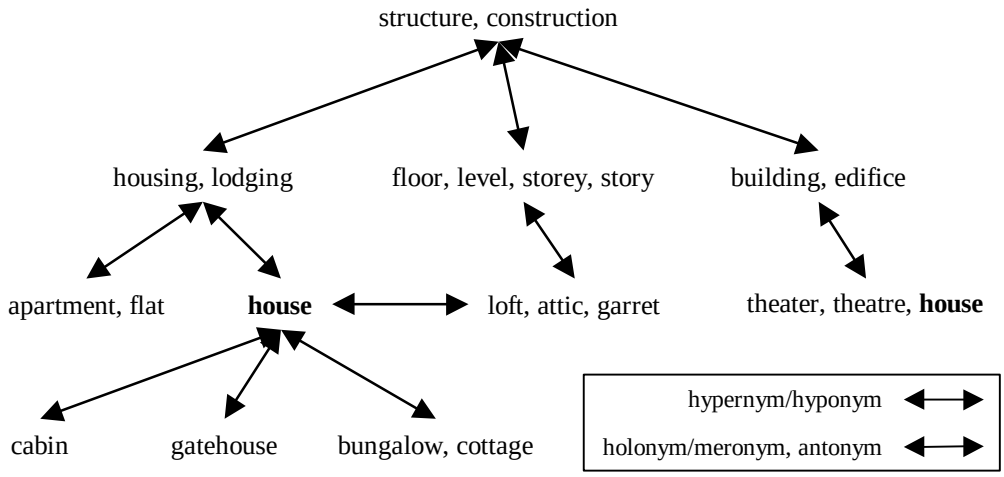
Similar to the issue in IR of human disagreement over judging the relevance of a document to a query, disambiguation researchers also face the problem of disagreements between people over what sense (or senses) a particular word occurrence is being used in. This issue was illustrated in a study performed by Kilgarriff *[Kilgarriff 91]* who, using a set of 83 words, tried to assign a single LDOCE sense to each word as it appeared in a number of contexts. Kilgarriff judged that 60 of the words had at least one occurrence that was not described by a single LDOCE sense. Such omissions place an upper bound on the measurable levels of disambiguation accuracy. Kilgarriff, however, did not discuss what this limit is. A debate between researchers on this matter is ongoing, *[Kilgarriff 97, Wilks 97]*.

3.2.6 Summing up

As illustrated with the research presented here, word sense disambiguators moved from small rule-based systems with limited scope to become tools that could potentially be applied to large-scale problems. It was not clear, however, if their accuracy had necessarily been improved at the same time. It was in this context that a number of researchers attempted to apply these tools to whole document collections in an attempt to improve retrieval effectiveness.

4 Disambiguation and IR experiments

One of the first mentions of disambiguation and IR is the work of Weiss *[Weiss 73]* who described using a disambiguator to try to improve the representation of a document collection. He reported that experiments using an IR system had shown that resolving all ambiguous words in a document collection would only result in a 1% improvement in retrieval effectiveness. As Weiss did not describe his experiment in detail, one can only speculate on the reasons behind this claim. The first large scale experiments where a corpus-based disambiguator was applied to a document collection for subsequent use by an IR system, were the work of Voorhees, of Wallis and of Sussna. Their work is now described followed by a discussion on their results.



4.1 Voorhees

Voorhees *[Voorhees 93]* built a sense disambiguator based on WordNet. In order to understand the design of her disambiguator, is it necessary to know something about the structure of this thesaurus. Each of WordNet's 90,000 words and phrases is assigned to one or more *synsets.* A synset is a set of words that are synonyms of each other; the words of a synset defines it and its meaning. The synsets, to which a particular word is assigned, constitute the individual senses of that word. All synsets are linked together to form a mostly hierarchical semantic network based on the hypernymy (is-a) and hyponymy (instance-of) word relations, with some additional relations of meronymy (has-part/has-member), holonymy (member-of/part-of relation), and antonymy (is-opposite). Voorhees' disambiguator used just the synsets of nouns and the hypernym and hyponym hierarchical relations between them.

To disambiguate an ambiguous word occurrence, the synsets (i.e. senses) of that word were ranked based on the amount of co-occurrence between the word's context and words in the *hood* of its synsets. Voorhees defined the hood of a word sense contained in a synset *s* as follows.

> "To define the hood of a given synset, *s*, consider the set of synsets and the hyponymy & hypernymy relations in WordNet as the set of vertices and directed edges of a graph. Then the hood of a given synset *s* is the largest connected sub graph that contains *s*, contains only descendants of an ancestor of *s*, and contains no synset that has a descendent that includes another instance of a member [i.e. word] of *s*."

From the fragment of WordNet structure shown in the figure above, the hood of the first sense of "house" (a lodging) would include the words

> housing, lodging, apartment, flat, cabin, gatehouse, bungalow, cottage.

The words "structure" and "construction" (at the top of the structure) would not be included since one of the descendants of that synset contains another sense of the word "house".

Voorhees compared retrieval effectiveness on a disambiguated test collection (test collection queries were also automatically disambiguated) against the effectiveness on that collection in its original ambiguous state. The collections she ran these tests on were CACM *[Salton 83]*, CISI, Cranfield 1400, MEDLINE, and Time (see *[Virginia disc 90]*, *[Sparck Jones 76]* for information on these collections). The results of her experiments showed that for each of these collections, retrieval effectiveness was found to be consistently worse when retrieving from the sense tagged collections. Voorhees did not measure the accuracy of her disambiguator but instead conducted what she called a "subjective evaluation" of it. Her conclusion from this was that it was tagging senses inaccurately and that this was the most likely cause of the poor results. In addition, for a number of the shorter test collection queries, she found the disambiguator was unable to accurately determine the intended sense of words within those queries.

4.2 Wallis

Wallis *[Wallis 93]* used a disambiguator, based on the work of Wilks et al. *[Wilks 90]*, as part of a more elaborate experiment to represent the words of a document collection by the text of their dictionary definitions. This was done so that synonymous words, which it was hoped would have similar dictionary definitions, would be represented in a similar manner and therefore documents containing synonymous words would be retrieved together. In his paper, Wallis illustrated this representation method using the synonymous words "ocean" and "sea":

> ocean (n) The *great* mass of *salt water* that *covers* most of the *earth*;
>
> sea (n) The *great* body of *salty water* that *covers* much of the *earth*'s surface.

As can be seen, these words do indeed have similar definitions. In choosing Wilks's disambiguation method, Wallis had the advantage of using LDOCE which, as was mentioned above, has its sense definitions written from a vocabulary of around 2,000 words. Wallis hoped this would result in a greater overlap in the definitions of synonymous words. When replacing a word occurrence by the text of its definition, if the word was ambiguous, the disambiguator was used to (hopefully) select the correct definition.

Wallis performed a number of tests on the CACM and Time collections. He measured the impact of his disambiguator on retrieval effectiveness and found a large drop when using a sense-based representation. Using this representation as a baseline, he then expanded the word senses with the LDOCE definitions and found no change in effectiveness. Like Voorhees, Wallis did not explicitly measure the accuracy of the disambiguator, so it is difficult to gauge what caused the drop in effectiveness.

More recently Richardson & Smeaton *[Richardson 95]* attempted a similar approach of replacing words in a document collection with a representation derived this time from the WordNet semantic network. Again a disambiguator was required to select correct senses. The results from their retrieval experiments were similarly disappointing. Later when Smeaton & Quigley *[Smeaton 96]* took the same approach of using an expanded sense based representation of a collection of manually disambiguated short image captions (5.9 words per caption on average), they found significant improvements in effectiveness. The very short length of the collection documents (the captions) undoubtedly contributed to the magnitude of the improvement however: any type of expansion on these sorts of documents will have been likely to be beneficial. Nevertheless, Smeaton & Quigley's work showed that using senses to help in term expansion is potentially a good strategy (see also Gonzalo et al's work below).

4.3 Sussna

Sussna *[Sussna 93]* expanded on Voorhees' WordNet based work using more of the thesaurus's semantic relations. His aim was to use the thesaurus to enable him to calculate a semantic distance between any two words. To achieve this, he assigned a weight to all the relations in WordNet. The strength of weight assigned reflected the semantic similarity expressed by that relation. For example, synonymy was assigned the highest weight, whereas antonym relations were assigned the lowest. The semantic distance between two synsets was calculated by summing the weights attached to the relations making up the shortest path between those two synsets. Sussna made no mention of the potentially expensive search required to find this path in the network.

The disambiguation method Sussna used was similar to the methods outlined in previous sections: given an ambiguous word appearing in a certain context, all the synsets (senses) containing that word were looked up in WordNet. Each synset was given a score calculated as the sum of semantic distances between the context words and that synset. The score was used to rank the synsets, with the top one being chosen as the sense of the ambiguous word occurrence. Sussna tried his disambiguation technique in a number of configurations. The main parameters he varied were the size of context used when disambiguating a word and the number of words disambiguated simultaneously. When disambiguating more than one word at a time, Sussna's technique examined every sense combination and unlike Cowie's simulated annealing method, encountered problems of the sense combinations increasing exponentially. Testing was performed on ten documents taken from the Time collection. Within these documents 319 ambiguous word occurrences were selected and manually disambiguated by Sussna. The disambiguator resolved these occurrences with an accuracy of 56%. It was reported that a context of 41 words produced the best disambiguation accuracy. In addition, simultaneously disambiguating words improved accuracy, but due to the number of sense combinations growing exponentially, the disambiguator was limited to processing concurrently no more that ten ambiguous words.

Sussna reported using his disambiguator in a small retrieval experiment involving the Time collection *[Sussna 97]* where documents were indexed by word senses. He reported disappointing results, however, never managing to produce better retrieval than a version of SMART indexing just words.

4.4 Analysis of word sense disambiguation & IR

The initial attempts at automatic disambiguation and IR produced negative results. Three papers analysing ambiguity and IR were published that provide clues to the reasons for this lack of success. They are by Krovetz & Croft (who's paper pre-dates almost all of the work discussed in the previous section), by Sanderson (the author of this paper) and Gonzalo et al.

4.4.1 Krovetz & Croft

As part of a wide-ranging paper on disambiguation and IR, Krovetz & Croft conducted a large-scale study of certain hypotheses on the relationship of relevance to sense matches and mismatches between query words and document words *[Krovetz 92]*. Using the CACM and Time test collections, Krovetz & Croft examined the ten highest ranked documents retrieved from each collection query. They examined the sense match between each query word and its occurrence in each retrieved document. Their aim was to find if a relationship existed between sense matches/mismatches and the relevance/non-relevance of the documents. Their results showed such a relationship, finding that sense mismatches were significantly more likely to occur in non-relevant documents. In addition, they found that word senses mismatches were relatively uncommon in the top ranked documents. Krovetz & Croft speculated that this was due to two reasons.

- First, the query word *collocation* effect, which can be explained through an example. If one were to enter the query "bank" into an IR system, it is just as likely to retrieve economic documents as it is geographic ones (assuming the balance of these two senses is equal in the collection being searched, see below). If, however, one entered the query "bank economic financial monetary fiscal" then, for top ranked documents, it is likely that many of the query words will collocate in those documents. It would be unlikely that an occurrence of "bank" in such a document would refer to the margin of a river. Therefore, collocation can reduce the effect of ambiguity in a query's individual words and consequently lessen the need for disambiguation.
- Second, the senses of many words have a *skewed frequency distribution*. Krovetz & Croft showed that the senses of a word do not occur in equal amounts and that often one sense of a word occurs more frequently than the others. With this in mind, Krovetz & Croft examined the CACM collection to find the proportion of query words which would be worth disambiguating: those having senses not appearing with a skewed distribution (skewed defined as one sense occurring over 80% of the time); and those who had a skewed distribution but the query used the word in a minority sense. They found that 15.6% of words fell into the first category and 8.8% into the second. This left 75.6% of test collection query words either being unambiguous or having skewed senses and the query using the majority sense. For these query words, disambiguation would be largely unnecessary as the majority of their occurrences in the documents were used in the desired sense.

Since they had performed manual disambiguation on the queries and the ten top ranked documents of both test collections, Krovetz & Croft examined the improvement in retrieval effectiveness caused by removing retrieved documents which contained sense mismatches. On the Time collection a 4% increase in average precision was found, but on the CACM collection the increase was 33%, although Krovetz & Croft put much of this large improvement down to better stemming while the sense matching was being performed. They concluded by suggesting a number of situations where disambiguation may prove useful: where the effects of collocation were less prevalent such as high recall searches; and where query words have senses with uniform distributions or were used in a minority sense.

4.4.2 Sanderson

Sanderson's analyses *[Sanderson 94]*, *[Sanderson 97]*, have echoed some of Krovetz & Croft's work and have also investigated the impact of erroneous disambiguation on IR effectiveness.

Sanderson employed a form of artificial ambiguity known as *pseudo-words*. Concatenating a number of words chosen randomly from a corpus forms a pseudo-word. These words become the *pseudo-senses* of a newly formed pseudo-word and all of their occurrences within that corpus are replaced by it: for example, randomly choosing the words "banana", "kalashnikov" & "anecdote", and replacing all their occurrences in a collection by the pseudo-word

"banana/kalashnikov/anecdote". A word can only be a member of one pseudo-word. By adding pseudo-words into a document test collection, a measurable amount of additional ambiguity is introduced into that collection and its impact on retrieval effectiveness can be determined. A pseudo-word with *n* senses is referred to as a size *n* pseudo-word.

Sanderson used the CACM, Cranfield 1400 and TREC-B test collections as well as the Reuters 22,173 text categorisation collection (created by Hayes *[Hayes 90]*, modified by Lewis *[Lewis 91]*) for his experiments. Unlike a conventional IR test collection, Reuters did not have a set of standard queries with a corresponding set of relevant documents. However, each document in Reuters was tagged with a number of manually assigned subject codes, which allowed Reuters to be used as a form of test collection. To do this, its documents were randomly partitioned into two equal sets: *q* the query set, and *t* the test set. The set *s* was defined as the set of all subject codes that were assigned to at least one document in both *q* and *t*. A retrieval was performed for each code in *s*; it was conducted as follows. Documents in *q* tagged with a particular code were selected and a 'query' was generated from them by term selection from relevance feedback. Retrieval was performed on the documents in set *t*. Effectiveness was measured by noting where in the ranking documents tagged with the subject code were (these documents were regarded as the relevant documents). An advantage of using this type of collection was that the size of the 'query' could be varied at will by altering the number of terms selected from relevance feedback.

Sanderson measured the effectiveness of an IR system retrieving from the collections and then measured it again after pseudo-words were introduced into the collection. The drop in effectiveness resulting from their introduction was a measure of the impact of that ambiguity. The results of the experiments showed that the introduced ambiguity did not reduce effectiveness as much as might have been expected. The analysis of the results concentrated on the length of queries showing that the effectiveness of retrievals based on a query of one or two words was greatly affected by the introduction of ambiguity but much less so for longer queries, confirming the collocation effect noted by Krovetz & Croft.

Although not stated in the paper, query term collocation within a document is also dependent on the document's length. If those being retrieved were particularly short (e.g. the image captions of Smeaton & Quigley's work *[Smeaton 96]*) then collocation of query terms, regardless of query size, is likely to be low. Therefore, in a situation of retrieving from short documents, one would expect to see the same impact of ambiguity on retrieval effectiveness as was observed with short queries.

Sanderson also used pseudo-words to study the impact of automatic disambiguation on retrieval effectiveness, concentrating particularly on the impact of disambiguation errors. By resolving the ambiguity in the pseudo-words[4] and occasionally making a mistake, he found that a 20%-30% error rate (the rate varied across the collections) could cause effectiveness to be as bad or even worse than when ambiguity was left unresolved. In other words, mistakes in disambiguation could be worse than the original ambiguity.

Sanderson concluded that a disambiguator was only of use in a retrieval context if queries were short or if disambiguation was performed at a high level of accuracy.

4.4.3 Gonzalo

Gonzalo et al *[Gonzalo 98]* transformed a manually disambiguated portion of the Brown corpus, known as SEMCOR, into an IR test collection to allow them to examine the benefits of retrieving from an accurately disambiguated document collection.

The test collection was created as follows. Gonzalo et al partitioned SEMCOR into 177 'documents'; a document was defined as a "coherent chunk of text". Next, a summary for each of the documents was written and manually sense tagged. On average the summaries were 22 words long. Each summary was regarded as a query. The document the summary was written for was regarded as the query's single relevant document; retrieval on a collection of this type is often referred to as *known item retrieval*. It was not stated who wrote the summaries or what they knew of the experiments.

[4] This was a trivial task since it was known which pseudo-sense (i.e. word) a pseudo-word occurrence was used in.

Rank	Words (%)	Senses (%)
1	48	53
5	84	84
10	91	89

Table 1: Gonzalo et al's results: percentage of relevant documents found at given rank.

For their experiments, Gonzalo et al first measured the effectiveness of their retrieval system when the queries and documents were represented by just words. Next, they retrieved using a sense-based representation and measured effectiveness again. Their presentation of results was an unconventional use of a recall/precision graph. But, it is possible to translate their figures into a more conventional presentation of *known item retrieval*: the percentage of queries where the known item was retrieved in the top 1, the top 5 and the top 10 ranked documents (see Table 1). Presented in this manner, the results show the sense representation was 11% better for items ranked at position 1. Further down the ranking, however, words and senses were equally good with perhaps words being slightly better.

One explanation for this was the likely presence of human error in the manual disambiguation of SEMCOR. However, Gonzalo et al claimed that it was due to part of speech differences between query and document words. They cited an example of a query using the word "design" in a verbal sense but the document using it as a noun. The example reveals a potential problem with disambiguation: words might be resolved to senses which are too specific for information retrieval purposes. It is possible that, for a query about newspapers, resolving down to a single sense (newspaper as a business concern as opposed to the physical object, for example) may be detrimental to retrieval. WordNet provides relationships that group senses (similar to LDOCE homographs). Such groupings, however, do not typically extend across grammatical categories, so the example highlighted by Gonzalo et al still remains. Krovetz has investigated ways of determining such relatedness from clues in dictionary definitions. His work is discussed more in Section 5.2.

Similar to the work of Wallis, of Smeaton & Richardson and of Smeaton & Quigley, Gonzalo et al also examined the utility of a representation based on WordNet synsets. The sense tags in SEMCOR were from WordNet, therefore, it was easy to translate a sense tag into its corresponding WordNet synset. Because synonyms of a word sense were part of the same synset, Gonzalo et al believed the representation would be richer. Retrieval based on synsets was found to to produce consistently high effectiveness results (62% of known items retrieved at rank one) better than words (48%) or senses alone (53.2%). Even when words in the documents were represented by all of their possible synsets (no disambiguation conducted), retrieval effectiveness increased (52.6%), an unusual result: such a simplistic approach would not normally be expected to work. Clearly the grouping of synonyms, which the synset representation provides, had a positive impact in this retrieval situation.

Like Sanderson, Gonzalo et al examined the issue of erroneous disambiguation and its impact on retrieval effectiveness. Working with the synset representation and using the no disambiguation case as a baseline (52.6%), the results showed that a disambiguation error of 30% is just higher than the baseline (54.4%) and an error rate of 60% is just below (49.1%). Although an error level similar to the baseline was not provided, one might speculate that such a level lies somewhere between 40-50%. This is lower than the levels stated by Sanderson (20-30%). As Gonzalo et al states, the difference between these results is likely due to the differences in sense representation: Sanderson using word senses; Gonzalo et al using synsets. It is also possible that the differences in the method used to measure retrieval effectiveness could be a factor: Sanderson using standard test collections; Gonzalo et al using known item retrieval. Repetition of their synset based work on a larger test collection with longer documents would be the logical next step.

4.4.4 In summary

These analyses seemed to point to three factors affecting the disambiguation and IR research presented so far.

- The skewed distribution of the senses of many words along with word collocation effects are responsible for the relatively small impact of ambiguity on retrieval effectiveness. In situations where these factors are not so prevalent, for example, if queries or documents are short, ambiguity is likely to be a problem.

- The accuracy of a disambiguator must be good for it to be of any benefit. There is no reason to suppose, that any of the researchers cited here managed to build a disambiguator better than Ng & Lee's which used extensive corpus based evidence and yet only managed an accuracy of 63.7%[5].
- A simple dictionary (or thesaurus) based word sense representation has not been shown to greatly improve retrieval effectiveness. This may be due to the inadequacy of single 'fine grained' senses as a means of representing documents for retrieval purposes: broader semantic groupings that cross grammatical boundaries, may be more appropriate. Alternatively, one may use a fine-grained sense representation as a starting point for other techniques such as synonym expansion.

4.5 Retrieval using ranked senses

Sanderson attempted to address some of these issues in his thesis work *[Sanderson 97]*. To this end, he tried to use disambiguator sense output in a more flexible manner and employed the variable length queries of the Reuters collection to focus on one retrieval situation where previous research had indicated disambiguation would be best suited: namely shorter queries. The disambiguator used was a version of Yarowsky's (see Section 3.2.4), modified to disambiguate WordNet senses. The disambiguator would, when given an occurrence of an ambiguous word, assign all the word's possible senses (including those in all grammatical variants) with a confidence score. Rather than represent the occurrence by the sense with the highest score (the usual method), Sanderson used a set representation composed of all the senses each with their assigned score. Similarity between two occurrences of a word was calculated by assessing the degree of similarity between the two sets. It was hoped that this representation would tackle some of the problems with single sense representations discussed in the Section above. In addition, it was also expected to be useful in handling the output of human disambiguation which sometimes included disagreements over a particular occurrence or situations where a person assigned more than one sense to an occurrence.

Unfortunately (again) the experimental results proved to be disappointing. Results showed that retrieval based on word senses, regardless of representation, produced worse effectiveness results than retrieval based on words alone. Disambiguator error was felt to be the cause of the poor results. There was an exception, however: retrieval from single word queries was better when documents and queries were represented by word senses. In addition, for this single positive result, Sanderson found that retrieval on documents represented by a full sense ranking was better than retrieval on the more classical representation where a single top ranked sense was used.

4.5.1 Summing up IR and disambiguation based on externally defined senses

With the exception of the one small positive result in Sanderson's work, all attempts, cited here, at automatic disambiguation and IR have failed to improve effectiveness. In considering ways forward for this research, one possibility is to try using the more accurate coarse-grained disambiguation strategies used by Yarowsky. Alternatively, it is possible that a targeting strategy of only using disambiguation for certain queries, based Krovetz & Croft's ideas (see Section 4.4.1), could be beneficial. The two issues of collocation and skewed frequency could drive this targeting. However, as the next section reveals, another approach may also be fruitful.

4.6 Disambiguation based on senses derived from corpora

In contrast to the work reported above which always assumed the presence of a pre-defined set of word sense definitions, there has been another approach to disambiguation in IR: one where only the corpus being retrieved from is used to build the disambiguator. Two researchers have taken forms of this approach; their work is now described followed by a discussion of the pros and cons of their technique.

4.6.1 Zernik

Zernik *[Zernik 91]* concluded that dictionaries would not provide a good source of sense definitions for IR purposes because their sense distinctions were too fine and were often based on grammatical rather than semantic criteria. (Although this conclusion seems to have been based on observations from a single dictionary.) He decided instead to

[5] Perhaps using the synset representation from Gonzalo et al, this level of accuracy may be good enough to produce a small increase in retrieval effectiveness although no one has reported such a result.

cluster word occurrences based on their contexts taken from the collection to be retrieved from. Once generated, Zernik would attempt to associate the clusters with some dictionary senses.

To generate clusters for an ambiguous word, Zernik's method extracted a ten-word context window surrounding every occurrence (in a corpus) of the word to be disambiguated. The resulting concordance was clustered based on three criteria: the context words; the grammatical category of the ambiguous word's occurrence; and its derivational morphology. Zernik assumed that this simplistic clustering would provide a reasonable separation between word senses. The method used to associate the clusters with dictionary senses isn't clear in the paper; it is assumed to be a manual process.

Zernik states he tested his disambiguator on twenty words. He only reports on what he called the best and the worst. The best word was "train" with the disambiguator identifying the locomotive sense from the activity sense with an accuracy of 95% (although Zernik does not provide a baseline accuracy to allow assessment of the quality of this result). Zernik attributed this success to the two senses having a grammatical distinction: one being a noun and the other a verb. Distinguishing between the two senses of the word "office", "to take office" and the physical space, proved too much for the disambiguator as it failed to make the distinction accurately.

Using these twenty disambiguated words within a corpus, Zernik performed a retrieval and examined the change in what he called *retrieval accuracy*: presumably some form of precision-based evaluation measure. When retrieving with a query composed of thirty words, Zernik reported no change in retrieval accuracy. For retrievals based on a one-word query, however, Zernik stated that "accuracy [was] improved by up to 50%".

With its heavy reliance on collection statistics and little use of a dictionary, Zernik's technique has similarites to that conducted by Schütze & Pedersen who dispensed with a formal definition of sense altogether.

4.6.2 Schütze & Pedersen

Schütze & Pedersen *[Schütze 95]* were the first to publish results showing a disambiguator working successfully with an IR system, reporting a 14% improvement in retrieval effectiveness. Somewhat like Zernik's disambiguator, theirs used only a corpus as evidence. For each word to be disambiguated in the corpus, the context of every occurrence of that word within the corpus was examined and similar contexts were clustered based on context words alone. For example, similar contexts of the word "ball" might include a social gathering and perhaps a number of different sports: tennis, football and cricket. For Schütze & Pedersen's disambiguator, each one of these similar contexts constituted an individual sense of the word, making them quite different from senses as defined in a dictionary. It is unlikely for example that a dictionary would distinguish between different types of the sporting sense of "ball". In addition, Schütze & Pedersen's disambiguator only identified frequent senses of a word; they stated that a similar context was only identified as a sense if it occurred more than fifty times in the corpus. So different are Schütze & Pedersen's senses from the classic definition of the word, that they are referred to here as *word uses*. The difference between uses and senses extends to their frequency distributions as well. The requirement that uses must occur at least fifty times eliminates the very infrequent and therefore makes the distribution of uses less skewed than senses. In addition, breaking up a word's commonest sense into a number of uses (e.g. the sporting sense of "ball"), has a similar effect of reducing any skew. Consequently, Krovetz & Croft's results on the percentages of words with skewed distributions of senses (Section 4.4.1) are likely to be different if measured on word uses.

Schütze & Pedersen's results came out of a relatively small retrieval experiment run on the TREC-1 category B collection: 25 topics (queries), numbers 51-75, were used. The number of queries which could be processed was limited due to the heavy computational load caused by clustering similar contexts. Disambiguation of a word occurrence was (as usual) the process of ranking the word's possible uses by a score based on the overlap between its context and the uses' contexts. The output of the disambiguator was used in a number of ways: representing a word occurrence by just the word (the base case), by the highest ranked word use, by the union of the *n* top ranked uses and by a combination of word and top ranked uses. They found the best retrieval results came from the final representation when the number of highest ranked uses was three.

There is one aspect of Schütze & Pedersen's approach that is potentially problematic, however. Their experimental set up gave them one advantage: the test collection they used had particularly large queries, often over 100 words in length. So large in fact that Schütze & Pedersen's disambiguator could identify the uses of query words automatically, no manual identification was required. This leaves open the question of how easily users could mark up their query with word uses given that uses were defined solely by their cluster of surrounding context words. For example, the context words of the tennis use of "ball" were likely to include the names of many tennis players. Unless users knew

who these people were, they might not have been able to deduce the meaning of that use. Nevertheless, given what is known about collocation effects on ambiguity, to produce such an improvement in effectiveness indicates that Schütze & Pedersen's word use identification is a successful strategy[6].

5 Retrieval without identifying senses

In contrast to the work reported so far, there has been a small amount of research in IR where word senses were explicitly taken into account but not explicitly identified through disambiguation.

5.1 Krovetz & Croft's sense weighting

Most IR systems employ *inverse document frequency* (IDF) weights which are computed from the collection-wide frequency of a word. Krovetz & Croft *[Krovetz 92]* suggested that calculation of this weight could be modified to take into account the number of senses a word has. They reasoned that a word with fewer senses would be better at discriminating between relevant and non-relevant documents than a word with many senses. Thererfore, they created a modified IDF weight to model this effect. They conducted tests on the CACM and Time collections using the LDOCE to provide information on the number of senses a word had. Comparing the retrieval effectiveness of their new weight against a standard IDF weight, their results were inconclusive. They found a small improvement on the Time collection and a drop in effectiveness on the CACM. Examining the CACM collection, they found that the drop in effectiveness was due to a number of factors:

- words being used in senses or in phrases not covered by the dictionary and thus being poorly modelled by the modified IDF weight;
- the presence of junk phrases in the CACM queries (e.g. "I want articles dealing with…") affecting the sense weighting.

There was a working assumption in the experiment that LDOCE provided an accurate count of the number of senses a word was actually used in and that the senses of each word occurred in equal quantities. Krovetz & Croft acknowledged that this assumption was probably flawed, but as the only solution would have been to manually disambiguate all occurrences of all occurrences of query words in the CACM and Time collections, the authors did not pursue this issue any further. The idea of sense based weighting nevertheless remains an enticing one where improvements may yet be obtainable.

5.2 Stemming

A stemming algorithm tries to reduce a word variant to its root form. One potential source of stemming error is when a variant of a word is used in one sense only, but its root form is more ambiguous. For example stemming "training" down to "train". A stemmer like the well known Porter stemmer *[Porter 80]* makes a number of these conflation errors by being 'over-zealous' in its removal of suffixes. Krovetz *[Krovetz 93]* described these errors as an issue of word sense. In his paper, he described the design of a different stemming algorithm, known as KSTEM, which tried to avoid erroneous stemming by using evidence from a dictionary, the ubiquitous LDOCE. KSTEM removed suffixes from a word variant piece by piece and after each removal, looked-up the reduced word form in the dictionary. Krovetz assumed that if the word was in LDOCE, this meant the variant had a different meaning from its root form and should not be stemmed further. He measured the accuracy of this assumption and found that 40% of variants could have been stemmed safely, 20% could not ("university" and "universe" for example) and for the final 40%, the correctness of stemming depended on the context of the query word. For example, if the query word "train" referred to a form of transport, any occurrences of "training" in a document should not have been stemmed. Despite the apparent weakness of his assumption, when Krovetz tested his stemmer on a range of test collections, he found it to improve retrieval effectiveness over the Porter stemmer in three out of four of the collections but not by a large amount. When taking into account KSTEM's further advantage of producing fully formed words as opposed to the truncated word forms from Porter, Krovetz concluded that his KSTEM system was the better stemmer.

---

[6] There are parallels between retrieval based on word uses and retrieval using *pseudo relevance feedback* based query expansion techniques such as Local Context Analysis (LCA) from Xu & Croft *[Xu 96]*. Although it has not been clearly demonstrated, it is quite possible that both techniques cause the same positive impact on retrieval effectiveness. Ballesteros & Croft *[Ballesteros 98]* have reported success in using LCA to overcome ambiguity problems in *cross language IR*.

Krovetz *[Krovetz 97]* explored more sophisticated methods to discover relatedness of word senses from LDOCE. Examining the definition of words, he investigated the utility of two techniques.

- First, finding in a word's definition one of its morphological or grammatical variants. For example, finding the word "liable" in the definition of the legal sense of the word "liability"; or finding the verb sense of "design" within the definition of the noun "design" would indicate relatedness between these pairs. Krovetz devised a method which automatically identified sense-pairs between morphological variants with 90.7% precision.
- Second, using word overlap between the definitions of two senses of a word to indicate their relatedness or lack there of. This proved to be as good as the first technique with up to 93% of detected sense pairs being related.

In his thesis *[Krovetz 95]*, Krovetz described the implementation and testing of these methods reporting that, in general, they provided a slight improvement in retrieval effectiveness.

5.2.1 Corpus based stemming

An attempt to improve on both Porter and KSTEM was recently reported. The underlying aim of the work was to avoid erroneous stemming this time relying on evidence derived from a corpus. Xu & Croft *[Xu 98]* built a stemmer that used Porter's suffix removal rules to reduce a variant to its root, but checked the 'correctness' of the reduction by examining the corpus to be retrieved from. Checking involved computing, over the entire corpus, co-occurrence statistics between a variant and its root. If the two were found to co-occur within 100 words of each other more often than would be expected by chance, the stem was allowed to proceed. The statistic used to measure co-occurrence was the Expected Mutual Information Measure (EMIM) (described in Chapter 3 of Van Rijsbergen's book *[Van Rijsbergen 79]*.

By requiring the variant and root to occur within 100 words of each other, Xu & Croft were in effect requiring that the words had similar contexts. As was shown in the review of disambiguation research, similar contexts are a reasonable indicator of words having the same sense. Examples of the stemmer's output indicated the success of this approach. Working on the Wall Street Journal collection (WSJ), the stemmer reduced "stocks" to "stock", something KSTEM would have prevented[7], and it did not stem "news" to "new", a mistake Porter makes. This reliance on contexts and ignoring of formal definitions of word senses also echoes Schütze & Pedersen's work on word uses and Xu & Croft's examples illustrate corpus specific stems. In addition to "stocks" and "stock" being allowed in this financial newspaper, "bonds" and "bond" were stemmed. Also the variant "gases" was not stemmed to "gas" because in WSJ these words were using in quite different ways. In tests on the WSJ and WEST test collections, Xu & Croft's stemmer was found to outperform both Porter and KSTEM in terms of retrieval effectiveness[8].

In terms of future work in stemming, currently, stemmers apply suffix reduction rules uniformly across a collection. One possible advance would be to examine the context sensitive stems that Krovetz noted in his analysis. It might be possible to create rules that only apply in certain contexts.

6 Conclusions

When Lesk's 1998 paper on dictionary based disambiguation was published, it caused a spurt of interest in the retrieval from word senses. These initial efforts showed quite clearly that exploiting disambiguation in IR was a harder problem than first thought. The analysis papers of Krovetz & Croft and of Sanderson, identified three reasons for this: word collocation, skewed frequency distribution of senses and inaccurate disambiguation. Disambiguation research has now moved into other areas:

- Gonzalo et al and Smeaton & Quigley used manually disambiguated test collections to allow them to measure the actual retrieval benefits gained from 'perfect' disambiguation;
- Krovetz has investigated stemming errors that can be avoided through detection of related senses;

[7] stocks (n); a wooden instrument on a post with holes for the neck and hands.

[8] In some ways, comparisons between the Porter stemmer and these more sophisticated stemmers is unfair. Speed and efficiency were more important criteria in Porter's design. Resources such as stemming dictionaries were deliberately avoided to increase the efficiency of the algorithm.

- Schütze & Pedersen produced an elegant, if computationally intensive, solution to the problem of disambiguating by ignoring the sense definitions of existing reference works.

As has been stated in this paper, it is believed that there are a number of possible avenues for sense based research to pursue in the future, not least the related area of cross language retrieval where translation issues require knowledge of sense. Whether these new approaches will bring the originally anticipated benefits remains to be seen.

7 Acknowledgements

This paper is based in part on the author's Ph.D. thesis supervised by C.J. van Rijsbergen. However, thanks to the Herculean efforts of one of the reviewers, the paper is much transformed and for this, the author is very grateful.

Much of the writing of this paper was done while the author was working at the Center for Intelligent Information Retrieval at the University of Massachusetts. While there, the work was supported by the National Science Foundation, Library of Congress and Department of Commerce under co-operative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

8 References

Ballesteros 98
L. Ballesteros & W.B. Croft (1997). Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval, in Proceedings of ACM SIGIR Conference, 20: 84-91.

Cowie 92
J. Cowie, J. Guthrie & L. Guthrie (1992). Lexical disambiguation using simulated annealing, in Proceedings of COLING Conference: 359-365.

Gale 92
W. Gale, K.W. Church, D. Yarowsky (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs, in Proceedings of the ACL, 30: 249-256.

Garside 87
R. Garside (1987). The CLAWS word tagging system, in The computational analysis of english: a corpus based approach, R. Garside, G. Leech, G. Sampson Eds., Longman: 30-41.

Gonzalo 98
J. Gonzalo, F. Verdejo, I. Chugur and J. Cigarran (1998). Indexing with WordNet synsets can improve Text Retrieval, Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, Montreal.

Grolier
Grolier Multimedia Encyclopædia CD-ROM. Grolier Interactive Inc., 90 Sherman Turnpike, Danbury, CT 06816, USA

Hayes 90
P. J. Hayes (1990). Intelligent high volume text processing using shallow, domain specific techniques, in Working Notes, AAAI Spring Symposium on Text-Based Intelligent Systems: 134-138.

Kelly 75
E. Kelly & P. Stone (1975). Computer recognition of english word senses, in North-Holland Publishing Co., Amsterdam.

Kilgarriff 91
A. Kilgarriff (1991). Corpus word usages and dictionary word senses: What is the match? An empirical study, in Proceedings of the 7th conference, UW Centre for the New OED & Text Research Using Corpora.

Kilgarriff 97
A. Kilgarriff (1997). I don't believe in word senses, in Computers and the Humanities, 31(2): 91-113.

Kirkpatrick 88
Roget's thesaurus of English words and phrases (1988). New ed. prepared by B. Kirkpatrick. Harmondsworth: Penguin

Krovetz 92
R. Krovetz & W.B. Croft (1992). Lexical Ambiguity and Information Retrieval, in ACM Transactions on Information Systems, 10(1).

Krovetz 93
R. Krovetz (1993). Viewing morphology as an inference process, in Proceedings of ACM SIGIR Conference, 16: 191-202.

Krovetz 95
R. Krovetz (1995). Word Sense Disambiguation for Large Text Databases, Ph.D. dissertation, Computer Science Department, University of Massachusetts, Amherst, MA, USA.

Krovetz 97
R. Krovetz (1997). Homonomy and Polysemy in Information Retrieval, in the Proceedings of the COLING/ACL '97 conference.

Lesk 88
M. Lesk (1988). "They said true things, but called them by wrong names" – vocabulary problems in retrieval systems, in Proc. 4th Annual Conference of the University of Waterloo Centre for the New OED.

Lewis 91
D.D. Lewis (1991). Representation and learning in information retrieval, in PhD Thesis, COINS Technical Report 91-93: Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003.

Longman 88
Longman dictionary of contemporary English, New edition, Longman

Miller 90
G.A. Miller (1990). WordNet: An on-line lexical database, International Journal of Lexicography, 3(4): 235-312.

Ng 96
H.T. Ng & H.B. Lee (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Examplar-Based Approach, in Proceedings of the ACL, 34: 40-47.

Porter 80
M.F. Porter (1980). An algorithm for suffix stripping, in Program - automated library and information systems, 14(3): 130-137.

Richardson 95
R. Richardson & A.F. Smeaton (1995). Using WordNet in a knowledge-based approach to information retrieval, in Dublin City University Technical Report, (CA-0395).

Salton 83
G. Salton, E.A. Fox, H. Wu (1983). Extended Boolean Information Retrieval, in Communications of the ACM, 26(11): 1022-1036.

Sanderson 94
M. Sanderson (1994). Word sense disambiguation and information retrieval, in Proceedings of ACM SIGIR Conference, 17: 142-151.

Sanderson 97
M. Sanderson (1997). Word Sense Disambiguation and Information Retrieval, PhD Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK, 1997

Schütze 95
H. Schütze & J.O. Pedersen (1995). Information retrieval based on word senses, in Proceedings of the Symposium on Document Analysis and Information Retrieval, 4: 161- 175.

Small 82
S. Small & C. Rieger (1982). Parsing and comprehending with word experts (a theory and its realisation), in Strategies for Natural Language Processing, W.G. Lehnert & M. H. Ringle, Eds., LEA: 89-148.

Smeaton 96
A.F. Smeaton & I. Quigley (1996). Experiments on Using Semantic Distances Between Words in Image Caption Retrieval, in Proceedings of ACM SIGIR Conference, 19: 174-180.

Sparck Jones 76
K. Sparck Jones & C.J. van Rijsbergen (1976). Progress in documentation, in Journal of Documentation, 32(1): 59-75.

Sussna 93
M. Sussna (1993). Word sense disambiguation for free-text indexing using a massive semantic network, in Proceedings of the International Conference on Information & Knowledge Management (CIKM), 2: 67-74.

Sussna 97
M. Sussna (1997). Text retrieval using inference in semantic metanetworks, Ph. D. thesis, University of California, San Diego, 1997

Van Rijsbergen 79
Information retrieval (second edition), published by Butterworths, ASIN: 0408709510, 1979.

Virginia disc 90
The Virginia disc one CD-ROM, Virginia Polytechnic Institute & State University Press. Editor, Project Director, Principal Investigator E.A. Fox, Dept. of Computer Science 562 McBryde Hall, VPI&SU, Blacksburg, VA 24061-0106

Voorhees 93
E. M. Voorhees (1993). Using WordNet™ to disambiguate word sense for text retrieval, in Proceedings of ACM SIGIR Conference, (16): 171-180.

Wallis 93
P. Wallis (1993). Information retrieval based on paraphrase, in Proceedings of PACLING Conference, 1.

Weiss 73
S.F. Weiss (1973). Learning to disambiguate, in Information Storage and Retrieval, 9: 33-41.

Wilks 90
Y. Wilks, D. Fass, C. Guo, J.E. Mcdonald, T. Plate, B.M. Slator (1990). Providing Machine Tractable Dictionary Tools, in Machine Translation, 5: 99-154.

Wilks 97
Y. Wilks (1997). Senses and Texts, in Computers and the Humanities, 31(2).

WordNet
http://www.cogsci.princeton.edu/~wn/. WordNet was developed by the Cognitive Science Laboratory at Princeton University.

Xu 96
J. Xu & W.B. Croft. Query Expansion Using Local and Global Document Analysis, in Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR 96), pp. 4-11.

Xu 98
J. Xu & W.B. Croft (1998). Corpus-Based Stemming using Co-occurrence of Word Variants, in ACM Transactions on Information Systems, 16(1): 61-81.

Yarowsky 92
D. Yarowsky (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora, in Proceedings of COLING Conference: 454-460.

Zernik 91
U. Zernik (1991). TRAIN1 vs. TRAIN2: Tagging word senses in corpus, in Proceedings of RIAO 91, Intelligent Text and Image Handling: 567-585.