Word Sense Disambiguation in Information Retrieval Revisited

Christopher Stokoe
The University of Sunderland
Informatics Centre
St Peters Way
+44 (0)191 515 3291
christopher.stokoe@sund.ac.uk

Michael P. Oakes
The University of Sunderland
Informatics Centre
St Peters Way
+44 (0)191 515 3631
michael.oakes@sund.ac.uk

John Tait
The University of Sunderland
Informatics Centre
St Peters Way
+44 (0)191 515 2712
john.tait@sund.ac.uk

ABSTRACT

Word sense ambiguity is recognized as having a detrimental effect on the precision of information retrieval systems in general and web search systems in particular, due to the sparse nature of the queries involved. Despite continued research into the application of automated word sense disambiguation, the question remains as to whether less than 90% accurate automated word sense disambiguation can lead to improvements in retrieval effectiveness. In this study we explore the development and subsequent evaluation of a statistical word sense disambiguation system which demonstrates increased precision from a sense based vector space retrieval model over traditional TF*IDF techniques.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistics processing;* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval Models, Search Process.*

General Terms

Performance, Experimentation, Verification.

Keywords

Word Sense Disambiguation, Information Retrieval, Performance Evaluation

1. INTRODUCTION

Ambiguity in natural language has long been recognized as having a detrimental effect on the performance of text based information retrieval (IR) systems. Sometimes called the polysemy problem [6], the idea that a word form may have more than one meaning is entirely discounted in most traditional IR strategies. If only documents containing the relevant sense of a word in relation to a particular query were retrieved this would undoubtedly improve precision.

Over the past 12 years there has been a steady increase in the performance of computerized Word Sense Disambiguation (WSD) systems. However if we look at the most recent evaluation, SENSEVAL 2 [3], we note there is some way to go before the performance of these systems is comparable to the 96.8% accuracy that Gale, Church and Yarowski observed in humans [4]. A number of attempts to apply these techniques within IR have on the whole met with little success with the notable exception of Schütze and Pederson [16]. This has led several authors [14,4] to conclude that until such systems reach human accuracy their potential to provide performance benefits in IR are limited, due to inaccurate disambiguation confounding any potential improvements.

It is important to note that attempts to unite the fields of WSD and IR have not been reported particularly well in the literature. In many early works [21,22,19] there was little or no attempt to accurately evaluate the performance of the WSD in isolation. Where evaluation was carried out [16,9] it was over small unrepresentative samples of language and/or small IR test collections. Some of these problems were caused by the lack of available resources to evaluate Natural Language Engineering (NLE) systems. However, there were also problems with the small size and highly cohesive nature of the IR test collections at the time. The issue of resources has become less problematic over recent years as more manually disambiguated corpora and larger, more diverse, IR test collections have become available. This makes possible more rigorous evaluation of not only WSD but its potential to improve performance in IR.

In this study we investigate the use of a state of the art automated WSD system within a web IR framework. The focus of this research is to perform large scale evaluation of both the automated WSD algorithm and the IR system. Our aim is to demonstrate relative performance of an IR system using WSD compared to a baseline retrieval technique such as the vector space model.

2. RELATED WORK

Most of the early work relating to the integration of WSD into IR resulted in no improvement in precision. A more complete review of these systems can be found in the work of Sanderson [15]. These initial failures prompted a number of researchers [7,14,5] to examine ambiguity within IR collections in order to understand where the benefits of WSD might be found. From these a number of key works can be identified that directly contributed to decisions made during this study.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
Copyright 2003 ACM 1-58113-646-3/03/0007…$5.00.

Firstly Krovetz and Croft [7] used the CACM and Time collections to study the relationship between sense mismatches amongst query terms and their occurrences in the collection. They concluded that collocation and co-occurrence between query terms naturally performed some element of disambiguation. This, in conjunction with the work of Sanderson [14], has indicated that in domains where large numbers of terms in a query were common the potential for WSD to be of benefit was reduced. This conclusion has subsequently led to studies [15,2,1] examining WSD in web retrieval due to the short nature of these queries [17]. Krovetz and Croft also concluded that there existed a "Skewed Frequency Distribution" in these test collections where 75.6% of query terms were used in their most frequent sense 80% of the time. This is the strongest indication of the importance of frequency statistics to a potential WSD system and leads to the idea that benefits from disambiguation may not be found from the overall WSD accuracy but rather how successful your system is at disambiguating the rare cases where a word is used in an infrequent way.

Sanderson [14] used artificial pseudo-words [23] to attempt to measure the effects of ambiguity on the CACM, Cranfield, and TREC-B collections. Having introduced artificially ambiguous terms into these collections he measured the retrieval performance and evaluated the results against the baseline for the original collection. Sanderson found that queries consisting of "one or two terms" were heavily affected by ambiguity however over longer queries there was little measurable effect confirming the results of Krovetz & Croft [7]. Additionally, Sanderson used pseudo-words to analyze the effect of automated erroneous disambiguation on the collections. This work indicated that an error rate of between 20–30% was enough to negate any performance increase from resolving ambiguity. From this Sanderson concludes that improvements in IR effectiveness would be observed only if computational linguistics could provide disambiguation of above 90% accuracy. Although questions remain as to the validity of the work based on pseudo-words [23,5] it becomes clear that any use of automated WSD within IR needs to be undertaken with an eye to limiting the effect of erroneous disambiguation. Sanderson [15] returned to the problem of WSD and IR in 2000 when he offered three key factors that affect WSD for IR. Firstly, skewed distribution of senses and collocation query effects are the reason why ambiguity has only a small impact on IR performance. Secondly, in order to benefit from automated WSD you need highly accurate disambiguation. This statement is less precise than his 1994 conclusions. Finally, he concludes that simple dictionary or thesaurus based word representations have not been shown to offer improvements in IR and as such he advocates the use of broader semantic groupings.

The work of Schütze and Pederson [16] remains one of the clearest indications to date of the potential for WSD to improve the precision of an IR system. Their technique involved examining the context of every term in the TREC 1 category B collection and clustering them based entirely on the commonality of neighboring words. The idea behind this is that words used in a similar sense will share similar neighbors, and by building a vector spaced representation of this co-occurrence and identifying different directions in the model we can indicate different contexts. The "Word Uses" (contexts) that were derived from the corpus were extremely fine grained and based heavily on frequency due to the fact that contexts based on less than 50 observed uses were dropped. A token evaluation of the disambiguation was carried out over 10 words with performance averaging at 90% accuracy. Initial experimentation showed that when word sense rather than term was applied to a standard vector similarity model average precision for the standard 11 points of recall increased from 0.299 to 0.311 (An increase of 1.2% in absolute precision). Although their system showed positive results using strict disambiguation better results were demonstrated initially using a more fault tolerant approach and latterly through a combined word and sense model. When strict disambiguation was relaxed and a word occurrence was allowed to correspond to any of its 3 closest context vectors precision increased to 0.321, an increase of 7.4% relative to the word based model and an absolute increase of 2.2%. When they ranked their retrieval runs using the sum of a document's score from both the word and sense model performance increased 14.4 relative to words alone and absolute precision increased 4.3%. It is our belief that by relaxing the strict disambiguation and combining the word and sense based rankings their system managed to overcome some of the negative effects of erroneous disambiguation.

Finally Gonzalo et al [5] converted the manually sense tagged Semcor1.6 corpus into an IR test collection to evaluate retrieval from a gold standard disambiguated corpus. They performed a series of known item retrieval tasks using document summaries (avg. 22 Terms in length) as queries. The results of this work demonstrate an 11% increase in performance using the sense data contained in Semcor over a purely term based model. Gonzalo et al then examined the effects of erroneous disambiguation. Using the term based model as a baseline (52.6% accurate) they then simulated disambiguation at 70% accuracy and 40% accuracy. Results indicated that 70% accuracy was enough to increase retrieval performance by 2.2% whilst performance decreased 3.5% with a retrieval accuracy of 40%. Sanderson [15] later extrapolated a break-even point of 50 - 60% accurate disambiguation for this work which is significantly lower than the 70%-80% indicated by his earlier pseudo-word based experiments. Gonzalo et al explained this as being a result of the difference in sense representation used in the two experiments. This work offers indications that in certain retrieval tasks less than gold standard accuracy may yield performance increases. Given the range of 50 - 60% accuracy established as a breakeven point, contrasted with the performance of the top systems at SENSEVAL 2 [3], we note that state of the art all-words disambiguation has begun to reach the appropriate levels.

3. EXPERIMENTAL SETUP

All of the retrieval experiments in our study were conducted using the TREC WT10G [20] corpus. This corpus consists of 1.69 million web documents for which there are two available sets of 50 relevance judged queries. The relevance judgments were created using pooling, with the retrieved document sets from each system submitted to the TREC evaluation being assessed by NIST analysts in relation to each query. Our experimentation utilized the TREC9 Ad-hoc Retrieval queries, NIST ID 451 – 500, an example query (Topic: 468) is shown in Figure 1. The title tag indicates the exact terms used by our retrieval systems whilst the description and narrative indicate the underlying information need that relevant documents should address. Our decision to evaluate WSD within the framework of web retrieval was based on the evidence discussed (section 2) which indicated that benefits from

<num> Number: 468

<title> incandescent light bulb

<desc> Description:

Find documents that address the history of the incandescent light bulb.

<narr> Narrative:

A relevant document must provide information on who worked on the development of the incandescent light bulb. Relevant documents should include locations and dates of the development efforts. Documents that discuss unsuccessful development attempts and non-commercial use of incandescent light bulbs are considered relevant.

Figure 1. An example of a TREC 9 Web Track query

automated WSD were most likely to be seen in problem domains that use short queries.

Our disambiguation system was trained and evaluated using Semcor1.6 [8] which is distributed with WordNet [10], a thesaurus created at Princeton University. WordNet consists of 90,000 terms and collocates organized into Synsets. Each Synset contains words which are synonymous with each other, while the links between Synsets represent hypernymy and hyponomy relationships to form a hierarchical semantic network. Semcor is a manually sense tagged subset of the Brown Corpus consisting of 352 Documents split into three data sets (see Table 1). The tag set used in Semcor consists of the unique sense identifiers used within WordNet.

Table 1. A breakdown of the composition of the Semantic Concordance (Semcor) Distributed with Wordnet1.6.

	No. of Documents	No. of Words	No. of Sense Tagged Words
Brown1	103	198796	106724
Brown2	83	160936	86412
BrownV	166	316814	41525

4. METHODOLOGY

From the related works discussed in section 2 four clear ideas emerge:

1. Skewed frequency distributions coupled with the query term co-occurrence effect are the reasons why traditional IR techniques that don't take sense into account are not penalized severely.
2. The impact of inaccurate fine grained WSD has an extreme negative effect on the performance of an IR system.
3. To achieve increases in performance, it is imperative to minimize the impact of the inaccurate disambiguation.
4. The need for 90% accurate disambiguation in order to see performance increases remains questionable.

In order to test these ideas we ran several IR experiments to compare term vector space techniques against the performance of a word sense model. Our WSD is carried out using an algorithm specifically designed to take into account the ideas discussed above.

4.1 Word Sense Disambiguation System

The word sense disambiguation algorithm we developed is based on popular ideas from the literature [15,9,23] with an emphasis on statistical co-occurrence and collocation. Given the skewed frequency distribution effects observed in prior experimentation (section 2), our goal was to construct an algorithm that took this phenomenon into account.

In order to provide empirical knowledge for use by our WSD system we created a bootstrapped representation of the Brown1 document set which is part of the Semcor corpus. For each unique word sense in the collection we automatically captured immediately adjacent collocates, words that frequently co-occurred within a one sentence window, lemmas, and part_of_speech (POS) information (although this wasn't ultimately used). An example of the type of data gathered can be seen in Table 2. We chose to focus on co-occurrence and collocation due to previous studies that had indicated the high precision of disambiguation systems using these specific knowledge resources.

Table 2. Example of the information captured from Semcor

Sense_Tag	POS	Lemma	Co-occurrence	Collocates
agent %1:17:00	NN	agent agents	biological delivery dose enemy epidemic immunity	infectious agent infectious agents biological agent causative agent
talk %2:23:01	VB	talk talked talking	time life made house family job drinking night armchair about	to talk never talked i talked not talking while talking parents talked talk with talking of just talking had talked

Although Brown1 contained a broad example of language (23,393 unique senses with an average of 4.6 occurrences per sense), it is clear that in a large scale IR experiment our training data would be too sparse to provide the coverage we required. In order to resolve this problem we relied on the sense frequency statistics contained in WordNet. These statistics represent a count of the number of times each unique sense of a word was observed in the lexical resources that were used to produce WordNet. Given Krovetz and Croft's results (section 2) it seemed reasonable to assume that if we had no specific indication of sense then simply returning the most frequent/common sense of the word would yield high accuracy.

We experimented with a number of disambiguation strategies, but we were unable to find a more effective technique than applying each of our knowledge sources (collocates, co-occurrence, and sense frequency) in a stepwise fashion. Using a context window consisting of the sentence surrounding the target word we would identify all possible senses of the word. We would then examine the surrounding sentence if it contained any collocates we had observed from Semcor, the word would be tagged with the corresponding sense. We would then do the same for co-occurrences, and finally if we had no specific sense data for a word or if no co-occurrences or collocates were observed in the context window, we would tag the word based on the frequency statistics in WordNet. In cases where WordNet contained no information relating to a specific term we would assign NO_TAG; these cases were often proper names.

Due to the sparse nature of web queries there was not enough information to provide a context window for our disambiguation algorithm. Given the results of Krovetz and Croft's work we tagged query terms based on frequency alone under the assumption we would achieve roughly 75% accuracy. In addition, this strategy meant that when our system attempted to disambiguate words in the IR test collection for which it had sparse training data it would tag them as being the same sense as the query term due to our reliance on frequency as a fall back technique. This acts to limit the impact of assigning sense based on weak assumptions by effectively making our disambiguation behave in the same way as traditional TF*IDF.

4.2 Information Retrieval System

The testing rig used for our retrieval experiments consisted of two inverted indexes of the TREC WT10G corpus. The first index is term based and the second uses stems, these were produced using an implementation of the Porter stemming algorithm [12]. We then produced retrieval runs using TF*IDF [13] ranking for both the term and stem based indexes. In order to evaluate the benefits of using our automated WSD system in an IR task we decided to contrast the performance of both the term and stem vector space models with that of sense based implementations (SF*IDF).

In total we produced four retrieval runs which can be categorized as follows:

1) Term Based – Traditional TF*IDF Ranking.
2) Sense Based(T) – Sense frequency (SF*IDF).
3) Stem Based – TF*IDF performed using stems.
4) Sense Based(S) – Sense frequency based on disambiguating all terms in the corpus which have the same stem as a query term (SF*IDF).

To produce the disambiguated runs we used the relevant documents identified by the corresponding baseline run. For example, all query term occurrences from the documents identified in the Term Based run were disambiguated and then re-ranked according to sense frequency to produce the Sense Based(T) run. For the Sense Based(S) run we disambiguated all terms in the documents identified in the Stem Based run which matched the stem of a query term. The motivation for our stem experiments was the idea that we could capture the desirable recall coverage of stem retrieval whilst using the lemmatizing effect of our disambiguation to increase the precision.

5. EVALUATION

Although the main objective of this study was to evaluate the performance of WSD in IR it was integral that we examined the accuracy of our disambiguation in isolation so that we could quantify its effects when used in our IR experiments. To provide a benchmark for the performance of our automated WSD system we used it to disambiguate the Brown2 part of Semcor. This was a fine grained evaluation where, unless our WSD system assigned the exact associated gold standard tag contained in Brown2 to a word instance, it was marked as wrong. In instances where our system assigned NO_TAG we marked that word instance as not having been attempted. Results for this evaluation are contained in section 6 and are reported in terms of precision. To provide a baseline by which to measure the effect of our disambiguation on a collection we used a technique put forward by Ng and Lee [11]. They advocated a baseline precision comparison between a raw sense frequency disambiguation run and a WSD algorithm in order to indicate relative performance. For our purposes this also has the advantage of giving a clear indication whether our technique is more accurate than simply ignoring sense within a corpus. Given the observation that sense tagging based only on frequency statistics treats each instance of a word as being the same sense, this is equivalent to the behavior of TF*IDF.

To evaluate the performance of our IR system we used the standard metrics of precision and recall. The top 1000 documents per topic for each of our 4 runs were evaluated against the relevance judgments available for the query set. For comparisons between runs we graphed average precision for 11 points of recall and additionally we calculated average precision over all queries. In order to better evaluate performance we also calculated precision at 9 standard document retrieval levels.

6. RESULTS

The overall performance of our WSD system was very positive. If we examine the precision graph (Figure 2) it shows that our disambiguation had an overall accuracy of 62.1%.

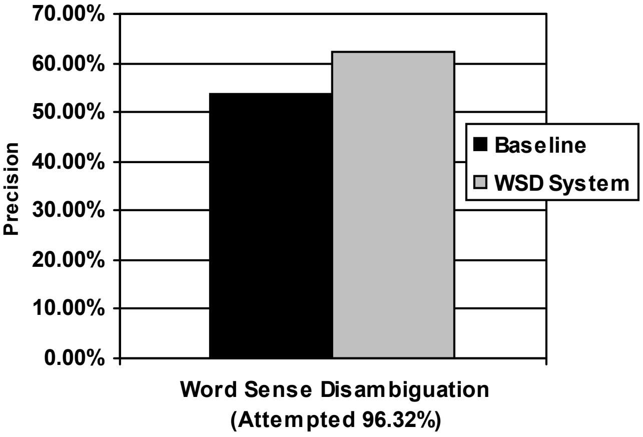

Figure 2. Comparison between the precision of our WSD algorithm compared to baseline frequency

Table 3. Results showing relative performance at 11 standard points of Recall

Recall	Term Based	Sense Based(T)	Stem Based	Sense Based(S)
at 0.00	0.2314	0.2082 -10	0.2444	0.2125 -13.1
at 0.10	0.0994	0.1318 +32.9	0.0922	0.1303 +41.3
at 0.20	0.0707	0.1061 +50	0.0587	0.0975 +66.1
at 0.30	0.0481	0.0872 +81.2	0.0396	0.0714 +80.3
at 0.40	0.0391	0.0627 +60.4	0.0356	0.0549 +54.2
at 0.50	0.0339	0.0447 +31.9	0.0314	0.0408 +29.9
at 0.60	0.0223	0.0298 +33.6	0.0247	0.0291 +17.8
at 0.70	0.0128	0.0184 +43.8	0.0112	0.0147 +31.3
at 0.80	0.0078	0.0115 +47.4	0.0046	0.0073 +58.7
at 0.90	0.0034	0.0065 +91.2	0.0036	0.0057 +58.3
at 1.00	0.0018	0.0045 +150	0.0012	0.0022 +83.3
Average Precision				
	0.0377	0.0550 +45.9	0.0340	0.0504 +48.2

Table 4. Table comparing Precision @ N docs for all 4 retrieval runs

Precision	Term Based	Sense Based(T)	Stem Based	Sense Based(S)
at 5 docs	0.0917	0.1125	0.1083	0.1042
at 10 docs	0.0875	0.1083	0.0833	0.1021
at 15 docs	0.0778	0.0889	0.0667	0.0833
at 20 docs	0.0677	0.0813	0.0625	0.0740
at 30 docs	0.0549	0.0688	0.0479	0.0667
at 100 docs	0.0292	0.0525	0.0231	0.0517
at 200 docs	0.0314	0.0489	0.0272	0.0474
at 500 docs	0.0307	0.0374	0.0297	0.0359
at 1000 docs	0.0240	0.0240	0.0230	0.0230
R-Precision				
	0.0425	0.0723 +70.1	0.0379	0.0627 +65.4

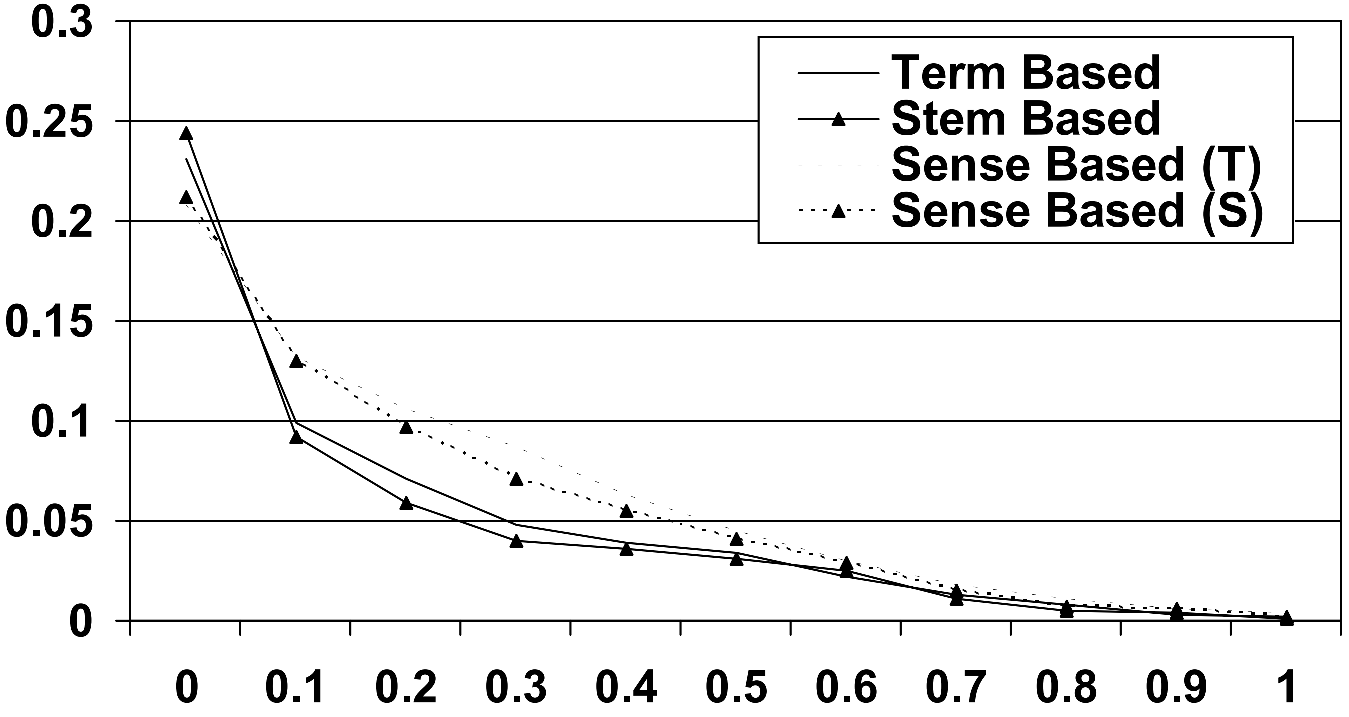

Figure 3. Plots precision for 11 recall points for the Term Based, Stem Based, Sense Based (T), and Sense Based (S) retrieval runs

This figure is above the breakeven point Sanderson identified from Gonzalo's work (section 2). However, it remains below the 70-80% indicated in his earlier study and indeed the 90% point he identified as being necessary for performance benefits to be achieved. In addition the algorithm attempted 96.32% of the 86,412 manually tagged examples in Brown2. This provides strong indicating that the coverage is sufficient for use on an IR test collection.

Figure 2 also shows a baseline frequency performance of 53.9% precision achieved by assigning word sense based on raw frequency statistics only. If we compare this with the performance of our WSD algorithm we see that it outperforms the baseline disambiguation by 8.2%. Given the fact that baseline frequency disambiguation effectively treats all term instances as being the same sense, which is in turn equivalent to the assumption made in traditional TF*IDF retrieval, we would expect an algorithm that outperformed baseline frequency to prove effective. This is dependant on the algorithm being used in such a way that it reduces the impact of erroneous disambiguation.

If we move on to examine the results of our retrieval experiments we observe a marked improvement in average precision when comparing both the term and stem runs to their automatically disambiguated equivalents. Table 3 shows average precision across 11 standard points of recall for all four retrieval runs as well as average precision over all relevant documents. When we compare the performance of the term model over the sense model we see that average precision jumps from 0.0377 to 0.0550. This is an increase of 1.73% in terms of absolute precision and a 45.9% increase relative to the performance of raw TF*IDF. When we look at the stem based results compared to its disambiguated equivalent we also see improved performance, with absolute precision increasing 1.64%, giving a relative performance increase over stem frequency of 48.2%.

By examining the precision / recall curve for all four runs (Figure 3) we can see that the main performance gains were in the mid to high recall range for both the term and stem models. However it is interesting to note that the WSD reduced the average precision in the low-recall range. Specifically both the word and stem based models showed a performance decrease at 0% interpolated recall. This was almost certainly a result of the known inherent instability observed at the low end of the recall curve. In fact, if we consider precision relative to the number of documents retrieved (Table 4), we note that at 5 documents the WSD model out performs raw TF*IDF by 2.1% in terms of absolute precision (a relative increase of 22.6%). However, this does not hold true when we consider the stem experiments where absolute precision drops 0.4% at 5 documents retrieved, a relative decrease of 3.8%. We do however note a subsequent increase in performance at 10 documents retrieved.

If we consider the results in terms of average R-Precision (Table 4), which is defined as the average precision of all queries @ N docs, where N is the number of relevant documents in relation to a specific query, we also see a clear improvement. When we compare the term and sense based model, we note that R-Precision increases from 0.0425 to 0.0723. This is an increase of 3% in absolute precision and a 70.1% increase relative to standard TF*IDF. Similar increases are also observed when we compare the stem model to its sense equivalent with performance increasing from 0.0379 to 0.0627, an absolute increase of 2.5% and a relative increase of 65.4%.

7. CONCLUSIONS

In this paper we have described a system that performs sense based information retrieval which, when used in a large scale IR experiment, demonstrated improved precision over the standard term based vector space model. Our disambiguation strategy used a combination of high precision techniques and sense frequency statistics in an attempt to reduce the impact of erroneous disambiguation on retrieval performance. Given the assumption that baseline frequency only disambiguation is in practical terms the equivalent of ignoring sense, it becomes clear that an automated disambiguation system should provide benefits in IR if it achieves higher precision than raw frequency. This however is not the case due to the profound negative effects of inaccurate disambiguation observed by both Sanderson and Gonzalo et al. In addition, it fails to take into account the resilience of traditional vector space techniques to the polysemy problem. This resilience is due to the skewed frequency distribution and query term co-occurrence effects observed by Krovetz and Croft. The success of our strategy lies in focusing on the high precision WSD techniques of collocation and co-occurrence whilst using raw sense frequency statistics to negate the low recall of such focused disambiguation. This in turn has the effect of capturing the positive performance of the WSD whilst in cases where we have sparse training data reducing the impact of the erroneous disambiguation to the baseline performance of TF*IDF.

If we examine this work in relation to Sanderson's 1994 claim (section 2) that less than 90% accurate disambiguation will not show performance increases in IR. We note that with an accuracy of only 62.1% our experimentation showed an absolute increase of 1.73% and a relative increase over TF*IDF of 45.9%. This certainly supports Gonzalo et al's less conservative claim that a breakeven point of 50-60% would be adequate. In addition Sanderson's skepticism as to whether simple dictionary/thesaurus sense definitions were adequate for use in this type of disambiguation also seems unfounded given that disambiguating terms into WordNet sense definitions proved effective within the scope of our work.

In the case of Schütze and Pederson, the only other experimentation to show significant performance increases, they achieved their best results through allowing a word to be tagged with up to three possible word senses and combining word and sense ranking. It seems clear that, as with our work, one of the key factors in the success of their experimentation was the steps they took to minimize the impact of erroneous disambiguation by introducing added tolerance. Given these observations three key ideas present themselves:

1. Less than gold standard disambiguation can provide increased precision in IR.
2. Once disambiguation accuracy moves past the performance of baseline frequency the problem becomes one of reducing the effect of erroneous disambiguation.
3. The benefits of using WSD in IR may be less than expected or only present within certain types of retrieval.

Although this work provides some interesting insights into the polysemy problem several areas of development remain. Firstly, this study is in no way a comparison of production systems, more an experimental evaluation of a sense based alternative to TF*IDF, a strategy which commonly forms an element of modern web document ranking. As such we ignored several key information sources that would be available to a full retrieval system such as document markup, link analysis, and similarity judgments. This is highlighted by the fact that the performance of our baseline model was significantly lower than the top systems in the TREC 9 evaluation. Secondly, our disambiguation strategy is, by certain standards very crude. The lack of context with short queries makes it extremely difficult to accurately disambiguate the query terms. We overcome this using frequency statistics however this leads to the potential for disastrous performance on queries where terms are used in an infrequent way. Thirdly, as with all attempts to use disambiguation within IR we rely heavily on the assumption that, there is a specific/correct way to interpret the query and that, the underlying information need is not in itself ambiguous. Finally, disambiguation which is heavily dependant on frequency may well prove inaccurate and nonproductive in natural language engineering endeavors such as machine translation. However, within the field of web IR where our goal is to improve on a simple bare term only model, the strategy yields significant performance increases.

8. FUTURE WORK

We are currently in the process of repeating this evaluation using the TREC 10 WT10G Q/REL set (NIST Topics 501 – 550), in order to provide further evidence to support our claims. Once this is completed we intend to carry out a topic by topic comparison of the term and sense based models in order to identify where the performance improvements were found. From this we hope to identify an optimal query length and/or degree of query polysemy associated with this technique. We also plan on expanding the training data for our disambiguation system to incorporate Brown2 in an attempt to increase the WSD accuracy.

In the long term, the key idea of engineering a WSD system and information retrieval mechanism in a manner that seeks to reduce the negative impact of inaccurate disambiguation merits further study. Although our solution worked through taking into account the skewed frequency effect observed by Krovetz and Croft, it is not as elegant as Schütze and Pederson's approach. Additionally, there is scope to explore the upper bounds for WSD performance within IR as disambiguation precision moves further beyond the baseline of sense frequency.

9. REFERENCES

[1] Agirre, E; Martinez, D. "Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web" Internal report: UPV-EHU, Donostia, Basque Country, 2000.

[2] Allan, J; Raghaven, H. "Using Part-of-speech Patterns to Reduce Query Ambiguity" In Proceedings of the 25th International ACM SIGIR, Pp 307 – 314. Tampere, Finland, 2002.

[3] Edmonds, P; Cotton, S. "SENSEVAL-2: Overview" In Proceedings of the Second International workshop on Evaluating Word Sense Disambiguation Systems. Toulouse, France, 2002.

[4] Gale, W; Church, K. W; Yarowski, D. "Estimating Upper and Lower Bounds on the Performance of Word Sense Disambiguation Programs" In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, Pp 249 – 256 Columbus, Ohio, 1992.

[5] Gonzalo, J; Verdejo, F; Chugur, I; Cigarran, J. "Indexing With WordNet Synsets Can Improve Text Retrieval" In proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Pp 38 – 44, Montreal, Canada, 1998.

[6] Kowalski, G; Maybury, M. "Information Storage and Retrieval Systems Theory and Implementation" Kluwer, Pp 97, 2000.

[7] Krovetz, R; Croft, W. B. "Lexical Ambiguity and Information Retrieval" in ACM Transactions on Information Retrieval Systems, Vol. 10(2), Pp 115 – 141, 1992.

[8] Landes, S; Leacock, C., Tengi, R. "Building Semantic Concordances" In WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.

[9] Li, H; Abe, N. "Word Clustering and Disambiguation Based on Co-occurrence Data." In proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Pp 749 – 755, Montreal, Canada, 1998.

[10] Miller, G. "Wordnet: an On-line Lexical Database" in Special Issue: International Journal of Lexicography Vol. 3(4). Pp 235 – 312, 1990.

[11] Ng, H.T; Lee, H.B. "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Examplar-Based Approach" in Proceedings of the 24th Annual Meeting of the Association for computational Linguistics, Pp 40 – 47, Columbia University, New York, 1996.

[12] Porter, M. F. "An Algorithm for Suffix Striping" Appeared in Readings in Information Retrieval, Ed. Spark-Jones, K; Willet, P, Morgan Kauffman, 1997.

[13] Salton G; McGill, M.J. "Introduction to Modern Information Retrieval" New York: McGraw & Hill, 1983.

[14] Sanderson, M. "Word Sense Disambiguation and Information Retrieval" In Proceedings of the 17th International ACM SIGIR, Pp 49 – 57, Dublin, IE, 1994.

[15] Sanderson, M. “Retrieving with Good Sense” In Information Retrieval, Vol. 2(1), Pp 49 – 69, 2000.

[16] Schütze, H; Pederson, J. O. “Information Retrieval Based on Word Senses” In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, Pp 161 – 175, Las Vegas, NV, 1995.

[17] Silverstein, C; Henzinger, M. “Analysis of a Very Large Altavista Query Log” SRC Technical note #1998-14, California, Digital Systems Reasearch Center: 17, 1998.

[18] Stevenson, M; Wilks, Y. “The Interaction of Knowledge Sources in Word Sense Disambiguation” Computational Linguistics, Vol. 27(3), Pp 321 – 349, 2001.

[19] Sussna, M. “Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network” In Proceedings of the 2nd International Conference on Information and Knowledge Management (CIKM), Pp 67 – 74, Washington, DC, 1993.

[20] Travis, B; Broader, A. “Web Search Quality vs. Informational Relevance” In Proceedings of the 2001 Infornortics Search Engines Meeting. Boston, 2001.

[21] Vooehees, E. M. “Using WordNet to Disambiguate Word Sense for Text Retrieval” In Proceedings of the 16th International ACM SIGIR Conference, Pp 171 – 180, Pittsburgh, PA, 1993.

[22] Wallis, P. “Information Retrieval Based on Paraphrase” In Proceedings of the 1st Pacific Association for Computational Linguistics Conference, Pp 118 – 126, Vancouver, 1993.

[23] Yarowsky, D. “One Sense Per Collocation” In Proceedings of the ARPA Human Language Technology Workshop, Pp 266 – 271, Princeton, NJ, 1993.