TREC 2002 – WEB TRACK
"Automated Word Sense Disambiguation for Internet Information Retrieval"

Christopher M. Stokoe
University of Sunderland
Informatics Centre
St Peters Campus
+44 (0)191 515 3291
Christopher.Stokoe@sund.ac.uk

Prof. John Tait
University of Sunderland
Informatics Centre
St Peters Campus
+44 (0)191 515 2712
John.Tait@sund.ac.uk

Abstract

We describe an attempt to use automated word sense disambiguation to improve the performance of an internet information retrieval system. A performance comparison of term frequency verses word sense frequency was carried out, the results of which indicated no significant performance gains from using a sense based retrieval model instead of the traditional TF*IDF.

1. Introduction.

Several authors have observed that ambiguity in language [1,2] can have a negative effect on the performance of text retrieval systems. Over the past ten years a number of researchers [1, 3] have worked on trying to integrate word sense disambiguation (WSD) techniques into text based information retrieval (IR) systems in an attempt to eliminate ambiguity and increase effectiveness. These attempts have on the whole produced disappointing results with the exception of Shütze and Pederson[4] who achieved a relative 14% increase in precision using a combined sense and term based model over traditional term only strategies. However it is important to note that often work in this field has been difficult to assess due to a failure to effectively evaluate the accuracy of the disambiguation used. Additionally there is evidence to suggest that often these experiments were undertaken on small or unsuitable collections. To this end the authors identified the need to re-examine the possible effects of automated word sense disambiguation in text retrieval systems using more rigorous performance measures.

2. Hypothesis.

Given that studies [5] have identified short queries may benefit most from a disambiguated collection we set out to evaluate the performance of automated word sense disambiguation within a web search system by submitting a class C entry to the TREC 2002 Web Track. The aim of this experimental work was to assess the relative benefits of searching from a collection with reduced ambiguity in an attempt to identify whether the introduction of automated word sense disambiguation can produce more effective results. In order to achieve this, the authors attempted to run a base line experiment taking effective performance measurements for both the disambiguation and retrieval models to assure the validity of the work.

3. Experiment Methodology.

To gain an accurate assessment of the effects of incorporating word sense disambiguation we created two full text indexes of the .GOV corpus. Each document in the collection was parsed using www::parser perl library and the resulting output was catalogued as plain text in *index(a)* and then fed into the automated word sense disambiguation system and sense tagged (see section 4). The sense tagged version was then added to *index(b)*. The format of the indexes used was relatively crude due to time constraints and although this subsequently effected retrieval times this was considered irrelevant in the scope of our experimental goals. Total index time was 7Days and 4Hours for *index(a)* and 22days 3Hours for *index(b)*. The indexing was carried out on a 1GHZ Pentium 3 with 398Mb of memory running Linux.

Once the indexes had been completed two topic distillation runs were performed and submitted to NIST. These runs were designed specifically to contrast the performance of the word and sense index models.

4. Automated Disambiguation.

The disambiguation strategy used was a statistical system trained using the Brown1 part of Semcor1.6 which is distributed with WordNet. Semcor consist of a subset of the Brown corpus manually disambiguated against the sense definitions contained in WordNet. The main language features used by our disambiguation system were sense frequency from both Semcor and WordNet, idioms and co-occurrence statistics observed from Semcor. These techniques were combined based on their individual accuracy to provide a hierarchy by which to select the appropriate sense. Each was applied using a context window consisting of the sentence incorporating the target word to be disambiguated.

The algorithm used to perform disambiguation was relatively simple using a stepwise approach to move through the techniques until either finding an appropriate match or falling back to sense frequency. Techniques were applied in descending order of their individual performance as determined in previous experimentation.

The performance of the disambiguation engine was measured in Precision and Recall and evaluated using the Brown2 part of Semcor1.6. Brown2 consists of 86,412 sense tagged word instances representing a traditional all-words test collection. Results were encouraging with the system scoring 60.1% precision and 57.9% recall. Coverage of the test corpus in terms of words attempted was 96.32%. Overall the system performed above the current baseline for disambiguation systems established from the Senseval-2 literature [6].

5. Retrieval Technique.

The retrieval mechanism used in both runs was a Boolean "AND" search with the results being ranked based on Salton and McGill's TF*IDF [7] measure summed across all terms in a query. The queries were stop worded and a rudimentary form of stemming was incorporated. The first run (TDtfidf) was a base line run carried out using *index(a)* in order to asses the relative performance of our combined retrieval and topic distillation technique. Our second run (TDwsdtfidf) was carried out using *index(b)* with TF*IDF calculated using sense occurrence rather that term frequency.

Overall performance in terms of speed of query execution was poor; however this was to be expected given the simplistic nature of our index strategy. Average processing time per query was 37.3 minuets for the term frequency model lowering to 34.1 minuets for sense frequency.

6. Topic Distillation.

Our strategy for topic distillation was relatively simplistic but because of time constraints it was impossible to carry out more traditional approaches such as link analysis or frequency distribution. As such our technique involved a post processing task carried out over a query's results to identify multiple hits / instances of pages from the same site. Once multiple hits from a single site had been identified the system reduced their URL's to the lowest common point of agreement where there existed a page in the document collection, this we refer to as the topic root. This page was then returned with an aggregate of the combined weighting score of its constituent elements. The multiple occurrences were stripped from the results and replaced with the topic root page ranked appropriately. The rational behind this strategy was the hypothesis that the arrangement of a site across the directory structure of a web server could be used to effectively assess the best point at which to enter the site for a given query.

7. Results.

We submitted two Topic Distillation runs for evaluation the first TDtfidf used term frequency and the second TDwsdtfidf used sense frequency. Both runs were identical in terms of the number of documents / number of relevant documents retrieved. However an examination of the systems results indicates subtle differences in the rankings. Table 1 shows the average Precision (non-interpolated) and R-Precision figures for both runs across all 49 queries.

Table 1: Combined results for runs.

Run Tag	R-Precision	Average Precision (Non-Interpolated)
TDtfidf	0.0451	0.0211
TDwsdtfidf	0.0454	0.0211

R-Precision shows a small increase in the performance of the sense model (TDwsdtfidf) when compared to the term model (TDtfidf).

Table 2 shows a comparison of the Interpolated Recall Precision Averages. When examining the Interpolated Recall – Precision figures we see improved precision in the low recall range when using the sense frequency model.

Table 2: Interpolated Recall - Precision

Interpolated Recall	TDtfidf	TDwsdtfidf
At 0.0	0.2941	0.2952
At 0.1	0.0751	0.0760
At 0.2	0.0180	0.0181
At 0.3	0.0040	0.0038
At 0.4	0.0008	0.0008
At 0.5	0.0008	0.0008
At 0.6	0.0000	0.0000
At 0.7	0.0000	0.0000
At 0.8	0.0000	0.0000
At 0.9	0.0000	0.0000
At 1.0	0.0000	0.0000

From this we can see that using sense information helped to promote a small number of key resources.

8. Conclusion.

The main aim of this project was to assess whether automated word sense disambiguation could be used to improve retrieval effectiveness. Although the use of automated disambiguation did lead to a small (0.0003%) increase in R-Precision this is considered statistically insignificant and as such the overall results were disappointing. There are several possible explanations for this.

- Firstly, the Topic Distillation strategy used was weak and in many cases missed the optimum page to return. The technique tended to reduce a key resource to the highest possible entry point of a particular site.
- Secondly, although our WSD strategy tested strongly in terms of overall accuracy it relied heavily on WordNET'S frequency statistics for words that had not been encountered in training our system. This meant that if no training data was available for a word all instances would be assigned the same sense which effectively failed to reduce any ambiguity from the corpus. Therefore increased training data could potentially lead to performance benefits.

Despite these problems it is important to note that many previous attempts to use automated disambiguation in IR have significantly reduced the performance of IR models such as TF*IDF. Although the performance gains we achieved were minimal the 39.9% error rate of our disambiguation methodology did not have a negative impact on retrieval performance. This runs contrary to the findings of Sanderson, 2000 [1] however further investigation is needed to assess exactly how much ambiguity was removed from the .GOV corpus.

9. References.

[1] Sanderson, M. 2000. "Retrieving with good sense" in Information Retrieval Vol 2, No 1 Pp 49 – 69.

[2] Kowalski, Gerald; Maybury, Mark, 2000. "Information Storage and Retrieval Systems Theory and Implementation." Kluwer, Pp 97.

[3] Vooehees, E. M. (1993). "Using WordNet to disambiguate word sense for text retrieval" Appeared in proceedings of ACM SIGIR Conference (16): Pp 171-180.

[4] Shütze, H; Pederson, J. O. 1995 "Information Retrieval Based on Word Senses" in Proceedings of the Symposium on Document Analysis and Information Retrieval 4 Pp 161 -175.

[5] Krovetz, R; Croft, W. B. 1992. "Lexical Ambiguity and Information Retrieval" in ACM Transactions on Information Retrieval Vol 10 Issue 1.

[6] Edmounds, P; Cotton, S. 2002 "SENSEVAL-2: Overview" in Proceedings of the Second International workshop on Evaluating Word Sense Disambiguation Systems.

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