Towards a Sense Based Document Representation for Internet Information Retrieval

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

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 word sense as an alternate text representation within an information retrieval system in order to enhance retrieval effectiveness. A performance comparison between a term and sense based system was carried out indicating increased retrieval effectiveness using a sense based representation. These increases come about by using a retrieval strategy designed to down rank documents containing query terms identified as being used in an infrequent sense.

1. Introduction.

Lexical ambiguity has long been considered as having a negative impact on the performance of information retrieval (IR) systems. Despite a number of studies [10,5,8,7,9] into ambiguity and IR to date only two have demonstrated significant performance increases. In the first, Shütze and Pederson [8] used the computationally expensive approach of clustering co-occurrences within the collection. Each of the clusters in which a given word was found was considered a unique "word use" with each word use arguably representing an individual sense of the word. The second study by Stokoe, Oakes and Tait [9] took advantage of the skewed frequency distribution in test collections observed by Krovetz and Croft[5] to create a sense based retrieval strategy that down-ranked documents which contained infrequent senses of a word. An additional property of this approach was that in cases of inaccurate disambiguation the technique degrades gracefully to at worst the baseline performance of a term model.

One perceived failing of both of these studies was their evaluation setting. Both showed a comparable performance increase when contrasting TF*IDF ranking with those achieved using sense frequency (SF*IDF). Although this is a first step to demonstrating the worth of a sense based document representation it is clear that most modern information systems rely on a combination of techniques to assign rank. This is most clearly demonstrated when we compare the performance of the Stokoe, Oakes and Tait work against the Web Track submissions for TREC 9. A baseline ranking produced using only TF*IDF was considerably below the average performance achieved by systems at the evaluation. Given this we must question whether the performance increases demonstrated by this technique are simply an artefact of the low performance of the baseline retrieval method.

2. Hypothesis.

Given that, at a low level we see a sense based representation outperforming a term based model, it is our belief that this increased performance can carry over to a modern web retrieval system. In order to demonstrate this we undertook to construct a "full featured" term based topic distillation system and to compare its performance against an identical system which used a sense based model. Our aims were as follows:

1) Produce a term based system with average or above performance.
2) Produce a corresponding sense based system.
3) Compare and contrast the performance of the term based vs. sense based system.

For our disambiguation we used the Sunderland University Disambiguation System (SUDS) running in the configuration described in Stokoe, Oakes and Tait for comparability. In terms of the

topic distillation techniques to be used we selected a number of common ranking algorithms that were utilised in the 2002 evaluation.

3. Experimental Methodology.

All the experimental work was carried out using a 1 GHz Pentium 3 with 398Mb of memory running Linux. All documents were striped of their headers and HTML tags and an initial term based inverted index of the .GOV collection was produced reducing the collection from 21GB to 5.3GB on disk. Total processing time for the production of this index was 11hrs 23 mins. The full text of each document in the collection was also made available for subsequent processing. For each query all the documents containing the query terms were identified from the index and were then subsequently ranked in accordance with our retrieval strategy. These same documents were subsequently disambiguated and re-ranked using a sense based representation.

4. Automated Disambiguation.

The disambiguation system we used (SUDS) is based on a statistical language model constructed from the manually sense tagged Brown1 part of the Semcor corpus. Briefly, it uses a statistical analysis of collocation, co-occurrence and occurrence frequency in order to assign sense. A more detailed explanation of SUDS can be found in Stokoe, Oakes and Tait. SUDS normally work's using a context window consisting of the sentence encapsulating the target word. However in cases where this information is not available E.g. queries, SUDS relies on occurrence frequency stats to perform sense tagging. Therefore one of the underlying assumptions behind SUDS use in IR is that query terms will rarely be seen as examples of a term being used in an infrequent sense.

SUDS overall accuracy is reported at 62.1% when evaluated using the Brown2 part of SemCor, this is representative of the current state of the art systems[2]. However more notably it outperforms bare frequency tagging by 8.2%. Given that frequency only tagging treat's all terms in the collection as being non-polysemous, and in turn represents the best possible accuracy you could achieve by ignoring sense. This indicates that SUDS can

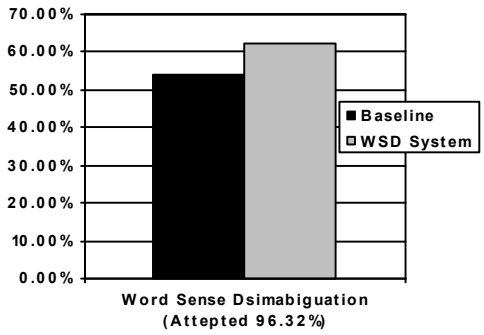

Figure 1: Comparison between the precision of our WSD algorithm compared to baseline frequency

provide a more accurate representation of a collection than simply ignoring sense given that it is more accurate than frequency only tagging.

5. Topic Distillation.

Our strategy for topic distillation was based on the increasingly popular link analysis theory initially proposed by Kleinberg [3]. This approach uses the notion that a key resource can either be an authority or a hub. Authorities are considered to be highly relevant documents of the type often in-linked by hubs which are inversely documents that contain significant out-links to authorities. Kleinberg proposed that by exploring the link topology of the WWW using a connectivity analysis mechanism one could make inferences on the relevance of a given document based on its linkages. Despite a number of key studies into the performance of link analysis as a ranking mechanism there remain some questions as to its effectiveness. In general the technique has demonstrated comparable performance to traditional statistical retrieval models. However in some cases reduced performance has occurred. These performance drops are generally perceived to be as a result of evaluation using a static collection with a high number of documents that contain out-links to documents not contained in the collection.

Given that traditional statistical ranking had performed reasonable favourably at TREC 2002 [1] we used a combination of ranking algorithms to identify authorities. The linkages between documents in our result set were then analysed in order to inflate the rankings of hubs by identifying those pages that had a significant number of out-links to other pages that were judged relevant by our system. Although we collected information about in-links to a given

document this was not used in our eventual ranking algorithm as we were unable to identify a way to use it which demonstrated increased retrieval effectiveness. Additionally our system tracked the number of pages form each unique domain which appeared in our final rankings allowing us to manipulate the number of results from each site that the system returned. This was eventually used in the "UNIQUE" runs in order to evaluate system performance where only the highest ranking page from a given site was returned. This was a common strategy at TREC 2002 where several groups [1] manipulated the number of unique sites that appeared in the top 10 retrieved results. Our sense based retrieval experiments followed exactly the same algorithm (see section 6) however terms were replaced by WordNet sense tags. Therefore each of our sense-based runs has a corresponding term-based baseline for comparison.

6. Retrieval Algorithm.

Our retrieval strategy used a number of common techniques associated with the vector space model of retrieval. Each query was stop worded to leave just the content terms and for each document which contained one or more of these terms we calculated an authority_rank using the following features:

1) TF*IDF

Using the well known ranking algorithm (1) presented by Salton and McGill[6]. With rank being assigned based on the sum of the weights of each term in the query.

$$W(w_i, d) = TF(w_i, d) * IDF(w_i) = N(W_{id}) * LOG\left(Nf / Nf(w_i)\right) \quad \textbf{(1)}$$

2) Cosine similarity (title)

A vector based comparison of the document title against the query. This was carried out using a cosine similarity measure (2). As seen in Salton and McGill [6].

$$\sigma(D, Q) = \frac{\sum_k (t_k \times q_k)}{\sqrt{\sum_k (t_k)^2} \times \sqrt{\sum_k (q_k)^2}} \quad \textbf{(2)}$$

3) Cosine similarity (body)

A vector based comparison of the document body and the query carried out using the similarity algorithm (2) we previously used on the document title.

4) Boolean Weighting

A weighting modifier applied based on testing whether a document is binary 'AND' complete for a given query. I.e. contains all of the content terms.

The rank assigned by each of these techniques was then normalised between 0..1 using max / min normalisation. Table 1 shows the weightings used in the max / min combination, added bias was given to documents that contained query terms in their <*title*></*title*> tags.

Feature	Weight
TF*IDF	1
TITLE_SIM	2
BODY_SIM	1
Boolean	1

Table 1: Weighting bias applied to each technique when merging the rankings using max / min combination.

Additionally for each document we calculated a hub_rank based on the sum of the authority_rank's assigned to any outward links contained in that document. In order to calculate the final page rank the hub_rank for a document was normalised and added to the authority_rank.

7. Results.

We submitted a total of four topic distillation runs for evaluation. Runtags beginning with SB indicate sense based runs while those beginning with TB indicate term based ones. Firstly if we examine the performance of the baseline term based runs (Table 2) we can see that TBBASE outperformed TBUNIQUE with regard to P@10. This demonstrates that returning only the highest ranked document from each website reduces overall system performance. Given that one key feature of topic distillation has always been to identify a suitable entry point to a relevant website it is interesting to note that our system gains performance when returning multiple pages from the same site. On several occasions our

Run Tag	R-Precision	Avg. Precision	P@10
TBBASE	0.1333	0.1166	0.1020
TBUNIQUE	0.1278	0.0948	0.0880
SBBASE	0.1283	0.1259	0.1020
SBUNIQUE	0.1407	0.1114	0.0940

Table 2: R-Precision, Avg. Precision, and Precision @ 10 for all runs.

TBBASE run demonstrates increased precision @10 by returning multiple ranked relevant pages from a single domain. If we consider the performance (precision @ 10) of our best baseline run compared with the median of all runs submitted to the evaluation (Figure 2) we can see that this run has average or above performance on all but five of the fifty topics. In addition we showed above average performance on thirteen of the topics.

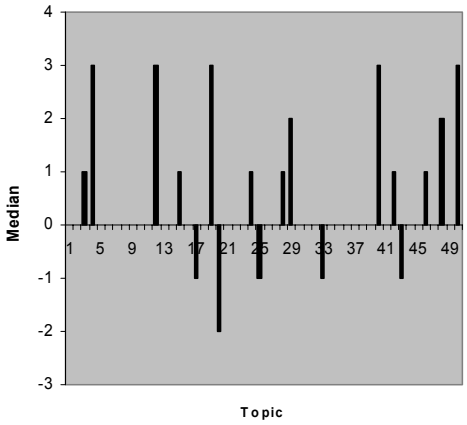

Figure 2: Difference from Median in Precision @ 10 per Topic.

Having established that the TBBASE run is representative of the current state of the art we can contrast its performance with the equivalent sense based run (SBBASE). The graph in Figure 3 shows the precision of both the TBBASE and SBBASE runs plotted for the 11 standard points of recall. We can see that the sense based run demonstrates increased precision compared with the term only model particularly in the mid-recall range. In addition if we compare the average precision of both runs 0.1166 (TBBASE) and 0.1259 (SBBASE) we note a small increase when using sense's rather than terms. There is an insignificant increase in Precision @ 5 (0.004) achieved using senses however the

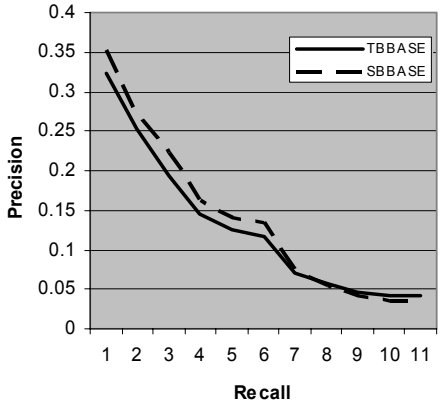

Figure 3: Precision – Recall for TBBASE and SBBASE @ 11 Standard Points of Recall.

term and sense models performed identically when we compare the official measure of precision @ 10.

8. Conclusion.

The main aim of our participation in TREC was to assess whether automated word sense disambiguation could be used to improve retrieval effectiveness. We developed a combination topic distillation system that used several traditional techniques and merged their rankings. A comparison between our term based system and the median of all runs in the Web Track topic distillation stream was performed. This demonstrated that our system was representative of average system performance levels seen at the evaluation. A comparison between the performance of our algorithm using a term and sense based representation was performed. The results of this evaluation demonstrate a clear performance gain from using sense information to represent documents. Although our sense based system failed to demonstrate increased precision @ 10 significant gains in recall were made without a corresponding drop in precision. In addition we do see increased avg. precision using a sense based representation and increased R-Precision. This increase was notable given that it was achieved using a disambiguation algorithm that has significantly lower levels of accuracy than those commonly associated with humans who perform the same task. One possible explanation for the success of this approach is that it exploits the skewed frequency distribution known to exist in large collections of natural language. The

work also offers anecdotal evidence to suggest there is a bias towards queries using polysemous terms in a frequently observed sense.

9. References.

[1] Craswell, N; Hawking, D. "Overview of the TREC 2002 Web Track" in proceedings of the Eleventh Text Retrieval Conference, TREC 2002.

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

[3] Kleinberg, J.M."Hubs, authorities, and communities" ACM Computing Surveys, 31(4es), 1999.

[4] Korfhage, R. "Information Storage and Retrieval" Wiley, New York, 1997.

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

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

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

[8] 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.

[9] Stokoe, C; Oakes, M. P; Tait, J. 2003 "Word sense disambiguation in information retrieval revisited." in proceedings of ACM SIGIR Conference (26) Pp 159-166

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