DUBLIN CITY UNIVERSITY
Ollscoil Chathair Bhaile Átha Cliath
Dublin City University, Glasnevin, Dublin 9, IRELAND.

# Using WordNet in a Knowledge-Based Approach to Information Retrieval

by

R. Richardson
A.F. Smeaton

School of Computer Applications
Working Paper: CA-0395

# Using WordNet in a Knowledge-Based Approach to Information Retrieval

Ray Richardson[1] and Alan F. Smeaton[2]
School of Computer Applications
Dublin City University
Glasnevin, Dublin 9, IRELAND

rrichardson@esri.ie, asmeaton@compapp.dcu.ie

**Abstract:** The application of natural language processing tools and techniques to information retrieval tasks has long since been identified as potentially useful for the quality of information retrieval. Traditionally, IR has been based on matching words or terms in a query with words or terms in a document. In this paper we introduce an approach to IR based on computing a semantic distance measurement between concepts or words and using this word distance to compute a similarity between a query and a document. Two such semantic distance measures are presented in this paper and both are benchmarked on queries and documents from the TREC collection. Although our results in terms of precision and recall are disappointing, we rationalise this in terms of our experimental setup and our results show promise for future work in this area.

## 1 Introduction

Many of the problems in information retrieval stem from the richness in terms of expressive power, yet the ambiguity inherent in natural language. Natural language ambiguity has long been recognised as a stumbling block to information processing in general. At the word level, humans have little difficulty in determining the intended sense of an ambiguous word for example, however without common sense knowledge we find it difficult to replicate this process computationally. Similarly, the multitude of ways in which the same concept can be described pose no trouble to humans but is a particular obstacle to successful information retrieval, (IR). Bates points out in [Bate86], "the probability of two persons using the same term in describing the same thing is less than 20%", and Furnas et al. found that '..the probability of two subjects picking the same term for a given entity ranged from 7% to 18%', [Furn87]. It is thus not surprising that only limited success is achievable with traditional IR approaches where information is viewed in terms of context independent single index and query terms matched as strings.

In our research we propose the use of a knowledge base (KB) as a controlled vocabulary and a semantic similarity function as the comparison mechanism between words. By anchoring intended senses of ambiguous terms in the knowledge base, some of the problems of natural language ambiguity can be addressed. Also by replacing direct string matching between index and query terms with a mechanism which can identify semantically similar terms, problems posed by the richness of natural language can be tackled.

Knowledge based information retrieval (KBIR) is not an entirely new concept. Previous work in this area includes the use of the Medline knowledge base as a controlled vocabulary in the Medlars medical IR system, [Rada88 and Rada89], the CoalSORT energy technology IR system

[1] The first author would like to acknowledge support from IBM IISL and Dublin City University.

[2] to whom correspondence should be addressed

[Mona87], and a system developed by Chen et al. which retrieved information specific to computing in the Eastern Bloc. However, a common characteristic of almost all KBIR systems to date is the fact they only operate in very specific and narrow domains. Perhaps the only exception to this can be found in Ginsberg's WorldViews system, [Gins93] though Ginsberg uses a manually constructed KB containing just 3,000 entries. It is very difficult to apply KBIR to the general domain with such a small KB. The objective of our research was to develop a domain independent KBIR system which used an automatically constructed KB containing an entry for all concepts found in everyday language. It was also our aim to propose sophisticated semantic similarity functions to operate within this KB. In the resulting KBIR system the KB was automatically constructed using the WordNet lexical database and two independent semantic similarity functions were derived.

The remainder of the paper is organised as follows. Section 2 describes how WordNet was adapted for use as a knowledge base. In section 3 there is a description of the similarity estimators we derived. Section 4 outlines how we tested these in an IR environment. The results of our experiments are presented in Section 5. In the final section we analyse why we got the results we did and we present conclusions and recommendations for future work in this area.

## 2. WordNet

WordNet is the product of a research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English, [Mill90a, Mill90b, and Beck92]. The system has the power of both an on-line thesaurus and an on-line dictionary, and much more, (refer to Figure 1). Information in WordNet is organised around logical groupings called synsets. Each synset consists of a list of synonymous word forms and semantic pointers that describe relationships between the current synset and other synsets. A word form can be a single word or two or more words connected by underscores, (referred to as collocations). The semantic pointers can be of a number of different types including :

- Hyponym/Hypernym (IS-A/ HAS A)
- Meronym/Holonym (Part-of / Has-Part)
- Meronym/Holonym (Member-of / Has-Member),
- Meronym/Holonym (Substance-of / Has-Substance)

In this work we only use the nouns from WordNet as a knowledge base, ignoring the verbs, adjectives and adverbs. The initial knowledge base consisted of a number hierarchical concept graphs, (HCGs), automatically constructed from WordNet data files. The root concepts of these HCGs were chosen as result of a set of experiments to determine what root concepts would, as a group, provide maximum coverage of the nouns in WordNet whilst minimising the degree of overlap between HCGs. The set of HCG roots we have used which achieves this is as follows :

- { Entity }
- { Location }
- { Abstraction }
- { Event }
- { Group }
- { Phenomenon }.
- { Psychological_feature }
- { Shape }
- { State }
- { Act }
- { Possession }

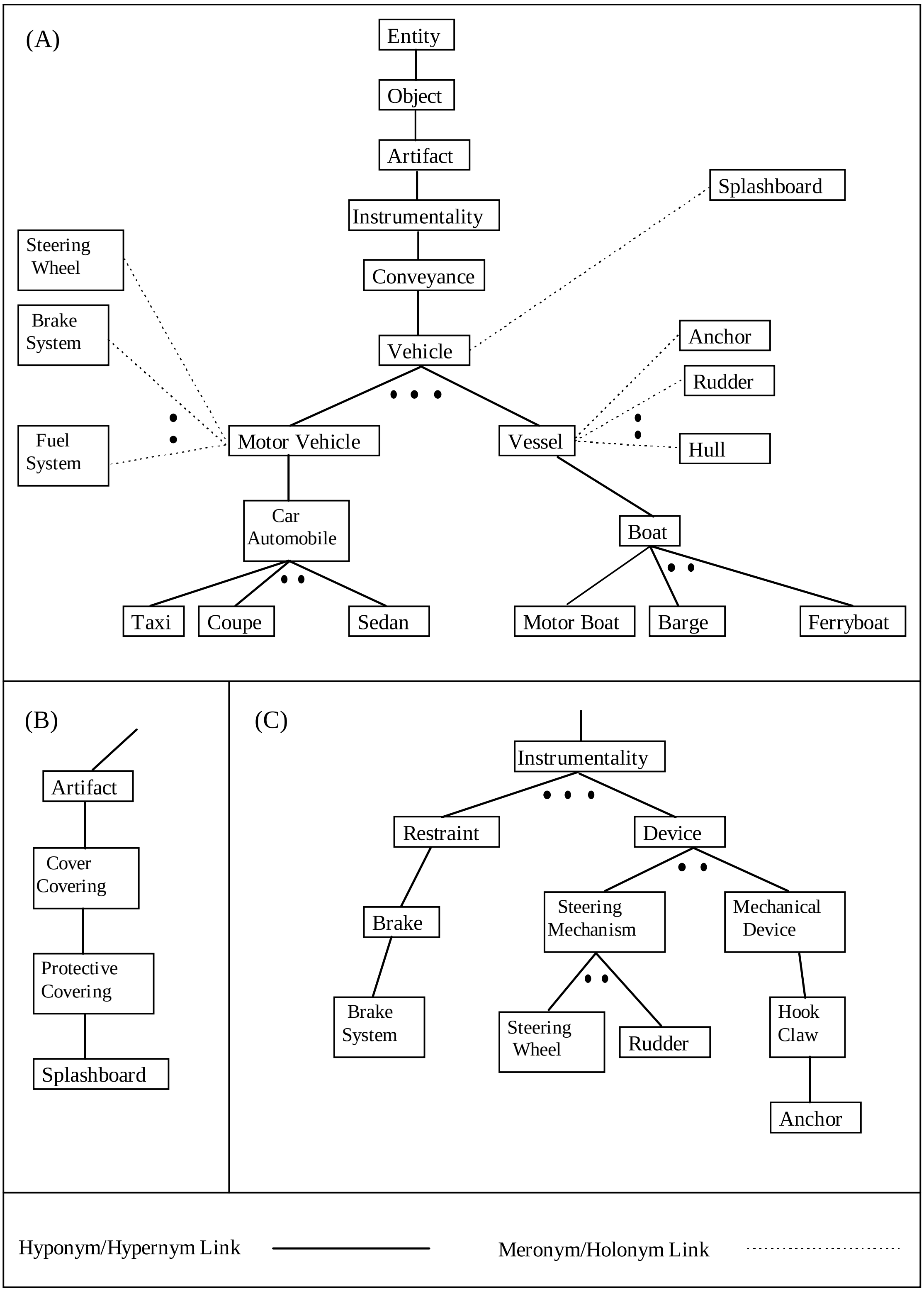

**Figure 1 : Extract from WordNet illustrating Lexical Inheritance**

The resulting HCGs ranged in size from 43950 unique concepts (Entity) to 688 concepts (Shape). The HCGs are organised in the same manner as the WordNet data files, being accessible via index files which index concepts by their byte offsets in the HCG file. One shortcoming of this simple and efficient organisation is that extending the files "is almost impossible", [Beck93].

Constructing the KB in this manner has its advantages and disadvantages. A significant advantage is the fact that the resulting HCGs will serve as comprehensive starting points in obtaining HCGs that contain all relevant concepts in an information domain. Also, WordNet based HCGs will contain a comparatively rich set of semantic link types.[3] However, foremost in the disadvantages is the fact that links in the resulting HCGs are not weighted. The following section addresses this problem more fully.

## 3. Word-Based Semantic Similarity

A number of approaches to measuring conceptual similarity between words have been taken in the past. Tversky's feature based similarity model, [Tvers77], is arguably the most powerful similarity model to date. However, its applicability in our IR situation would require a much richer knowledge base than is available to us. Although a WordNet-derived knowledge base is quite thorough in its coverage of concepts, the number of semantic relation types connecting these concepts is considerably less than would be required for use by a feature based similarity model.[4] As such, we experimented with two alternative approaches to estimating semantic similarity, and we refer to them as the information based and conceptual distance approaches.

### 3.1 The information based similarity estimator

The information based approach to measuring semantic similarity is based on work carried out by Resnick, [Resn93a, Resn93b]. Resnick views noun synsets as a class of words where the class is made up of all words in a synset as well as words in all directly or indirectly subordinate synsets. Conceptual similarity is considered in terms of class similarity. The similarity between two classes is approximated by the information content of the first class in the noun hierarchy that subsumes both classes. The information content of a class is approximated by estimating the probability of occurrence of the class in a large text corpus[5], ( see appendix A for a discussion on class probabilities). Thus the similarity of two classes can be expressed as :

$$Sim(c_1,c_2) = \max_{c_i}[\log \frac{1}{P(C_i)}] \qquad (1)$$

where {Ci} is the set of classes dominating both C1 and C2, P(Ci) is the class probability of class Ci, and $\log {}^{1}\!/_{P(C_i)}$ is the information content of class Ci.

The methodology could probably be best illustrated by example[6]. If we assume we wish to discover the similarities between the following classes : 'car', 'bicycle', 'banana', and 'fork'. Taking first Sim(car, bicycle), we see that WordNet has six classes to which both 'car' and 'bicycle' are subordinate :

[3] The concept graphs reported in [Rada89],[Kim90] and [Lee93] contain only IS-A links whilst the concepts in Ginsberg's WorldViews system, [Gins93], are only related by broader term and narrower term links.

[4] It is intended in the future to extend WordNet to include relation types of the form ATTRIBUTE-OF and FUNCTION-OF which will connect WordNet's adjective and verb collections with its noun collection. These developments should considerably enhance WordNet's applicability to feature based similarity models.

[5] In our case theprobabilities of occurrence of classes were computed from a collection of 11 million noun occurrences from the text of the Wall Street Journal

[6] This replicates one of Resnicks examples using version 1.4 of WordNet

| **Synset** | **Info_Content** |
|---|---|
| < vehicle > | 2.500 |
| <conveyance > | 2.433 |
| <instrumentality> | 1.338 |
| < artifact > | 0.980 |
| < object > | 0.763 |
| < entity > | 0.565 |

If one takes the similarity measure as being the  maximum information content value amongst the set of classes that subsume both synsets then SIM(car,bicycle) = 2.5.   Notice that, as would be expected, classes grow more frequent and as such less informative as one moves higher in the hierarchy.  Since 'car' and 'bicycle' have some specific (therefore informative) classes in common, one can conclude that they are similar.  In contrast, the other examples yield the following :

| **Sim(car,fork)** | | **Sim(car,banana)** | |
|---|---|---|---|
| <instrumentality> | 1.338 | < object > | 0.763 |
| < artifact > | 0.980 | < entity > | 0.565 |
| < object > | 0.763 | | |
| < entity > | 0.565 | | |

Thus we see that cars and forks seem quite a bit less similar than cars and bicycles, however they are more similar than cars and bananas.  This can be explained by the fact that forks and cars are objects that people make (artifacts), whereas all that can be said in terms of the similarity of cars and bananas is they are both nonliving things (object).

## 3.2 The conceptual distance estimator

The conceptual distance approach to estimating the semantic similarity between words is derived from work carried out by Rada.  In essence, the similarity between two concepts is estimated by the sum of edge weights along the shortest path connecting their corresponding synsets in the KB. This extends Rada’s definition of semantic similarity by including paths made up of meronym link types.

This estimator of semantic similarity assumes that the edges between synsets in the KB are weighted yet the relational links in WordNet are unweighted.  Also, unlike the concept graphs of others, ([Gins93], [Rada89], [Kim90], and [Lee93]), those created for our domain-independent KB are very large, containing of the order of tens of thousands of nodes.  For this reason, the usual process of hand weighting each link is not viable and a method of automatically weighting each link had to be developed.   Initial research in this area was based on Botafogo's work on node metrics in hierarchical hypertexts, [Bota92] however, our research was subsequently considerably influenced by that of Sussna, [Suss93].

Certain observations can be made with regard to conceptual distance in HCGs which can aid in the process of automatically determining the weight of edges.   For instance, the value for the weight of a link is affected by the density of the HCG at that point, the depth in the HCG and the strength of connotation between parent and child nodes.  With regard to the width, different parts of a HCG are denser than others.  For instance, the ’plant’ section of the KB is a very dense, (individual nodes having up to three and four hundred children), collection of generally unpronounceable plant species.  It can arguably be held that the distance between nodes in such a section of the concept graph should be very small, relative to other, less dense regions.  In terms of the depth it can be said that distance shrinks as one descends down a HCG.  To explain, suppose there are two only sibling relations, one near the top of the hierarchy and one deep down in the detailed portion of the HCG. To illustrate, suppose the node 'Living Thing' is high in the hierarchy and it only has two children nodes, 'Plant' and 'Animal'. These two siblings are far apart conceptually when compared against the two siblings 'Wolfhound' and Foxhound' under the parent 'Hound' deep down in the HCG. Finally, Figure 2 illustrates the point regarding the local strength of connotation.

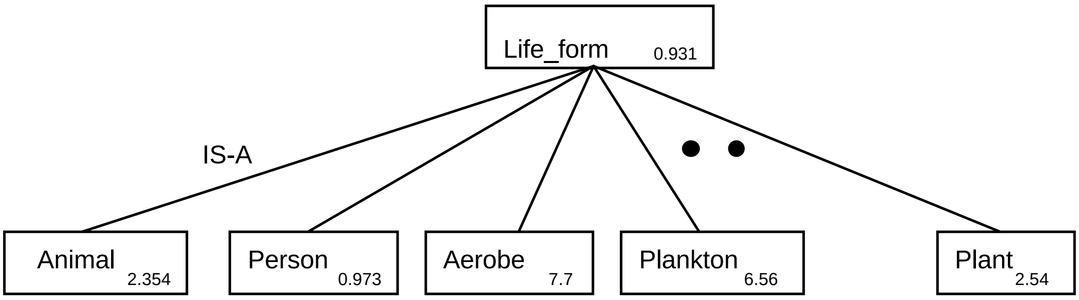

**Figure 2 : KB Extract**

It can be argued the parent node *Life_Form* is more strongly connotated with the child nodes *Animal*, *Plant*, and *Person*, than with the nodes *Aerobe* and *Plankton*.

At present, the density of a HCG for a specific link type is estimated by counting the number of links of that type. The strength of connotation of the link being weighted is estimated as a function of its information content value, and those of its sibling and parent node, (the numbers in Figure 2 are the nodes' information content values). The result of these two operations is then normalised by dividing by the depth of the link in the HCG.

### 3.3 Conclusions on our approaches to semantic similarity

Both estimators of semantic similarity are quite different in their approaches, and both have inherent strengths and weaknesses. Rada has shown through experiments that conceptual distance within a weighted HCG simulates, with surprising accuracy, humans in their assessment of the conceptual closeness between documents and queries. However, following some informal experimentation with the use of the conceptual distance measure, we found some general concerns with regard to the use of this measure as an estimator of semantic similarity. Due to the comparatively broad domain of our HCGs, (as compared with those of Rada who worked solely in the medical domain), the conceptual distance measures were less accurate than expected. The situation was improved to a large degree when it was decided to include the non-hierarchical link types in the distance calculation. However, the conceptual distance measure is still particularly susceptible to vagaries of the builders of WordNet. In particular, the organisation of concepts within WordNet can often be puzzling. The irregular densities of links between concepts results in unexpected conceptual distance measures. These are typically as a result of expected links between concepts not being present. Also due to the general operation of the conceptual distance similarity estimator most concepts in the middle to high sections of HCGs, being geographically close to each other, would therefore be deemed to be conceptually similar to each other. Although the depth scaling factor in the link weighting mechanism softens the overall effect in many cases, sometimes the general structure of the WordNet derived HCGs cannot be overcome by link weighting without causing serious side effects elsewhere in the HCG.

The information based measure of similarity is not as dependent on the existence and organisation of KB links as the conceptual distance measure. A certain amount of contextual information is captured from the text corpus used to calculate information content values, and this combined with the extensive coverage of concepts in our KB, provides us with a powerful measure of semantic similarity. This measure is still dependent on the organisation of concepts in the IS-A hierarchy, however, given the broad coverage of concepts in WordNet, it is difficult on the whole to be critical of the hierarchical structure of concepts. Also the authenticity of a synset's information content value is obviously dependent on the size and domain independence of the text corpus used. However, in our case, the use of 11 million noun occurrences from newspaper articles would seem to be a reasonable first attempt at calculating information content values.

Despite these apparent strengths of the information based similarity measure, it is not without weaknesses. Perhaps foremost is the fact that it ignores information in the KB that may be useful. Only the synonym and IS-A relations are used, the other relation types, which are used effectively by the conceptual distance approach, are overlooked. A second weakness is apparent in the method of calculating the information content of classes. Many polysemous words and multi-worded synsets will have an exaggerated information content value. If one takes for instance the word 'bank', the information content for this word will include all occurrences of bank in the corpus, regardless of word sense. This gives the same (exaggerated) information content value to a 'commercial bank' as a 'river bank'. Also, due to the fact information content values are calculated for synsets as opposed to individual words, it is possible for the information content value to be over-exaggerated in situations where synsets are made up of a number of commonly occurring ambiguous words. For example in the synset *{ yield, fruit }*, the information content value of this synset is calculated both from the frequencies of the word *'fruit'* and the word *'yield'*. Given the fact that the information content of a class is defined in terms of the information contents of its subordinate classes, super-classes of classes containing polysemous words are similarly over-valued. This disregard of ambiguous words is a problem given the fact that synsets in our WordNet-derived KB refer to particular senses of words and the KB as a whole tends to include very fine sense distinctions in an attempt to have an exhaustive coverage of concept meanings. A final caveat apparent with the information based approach to semantic similarity is the fact two different concepts can be more similar to each other than another concept is to itself. The effect of this can be more clearly seen with the following example :

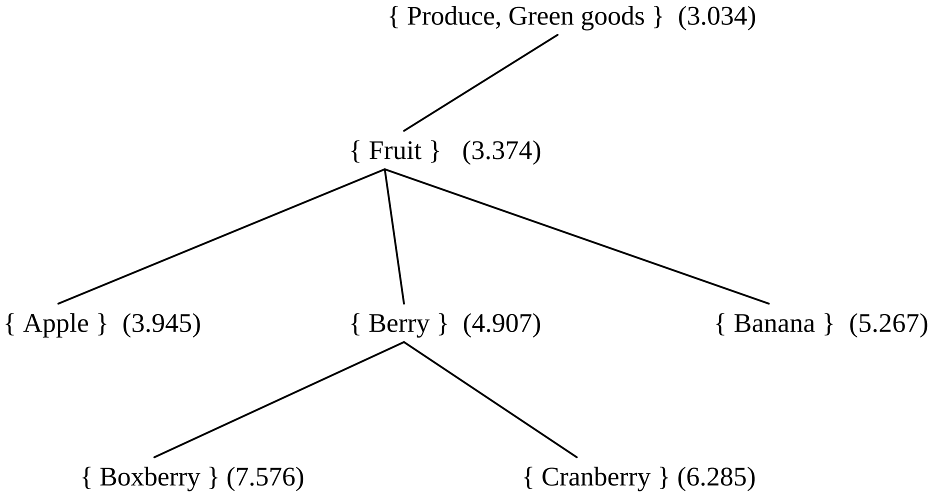

**Figure 3 KB Extract showing violation of minimality in information based similarity estimator**

From figure 3 we can see the information based estimate of the similarity between an apple and a banana (3.374) is closer than the estimated similarity between produce, (as in green goods), and itself (3.034). This is a clear violation of the minimality property of a metric. Violations of minimality and the other metric properties are undesirable in our simplified model of the world and could have a bad effect on system performance.

Despite there differences there was no obvious way of combining both similarity estimators to take advantages of the good aspects of each approach and compensate for individual weaknesses. Consequently we performed two sets of IR experiments, one using the conceptual distance similarity estimator and the other using the information based. Our objective was not just to determine which was the better measure but also to discover if one approach might be more suited to certain situations than the other and vice versa.

## 4. Evaluation

In this section we will briefly outline our experimental environment and implementation strategy. In section 4.1 there is a brief description of the TREC method of IR evaluation while an outline of how we implemented semantic-based document retrieval systems can be found in section 4.2.

### 4.1 Experimental Environment

The Text REtrieval Conference (TREC) is an annual benchmarking exercise among information retrieval researchers which is organised by NIST. In essence, many information retrieval systems run a number of the same queries or topics against the same document database, at the same period in time. The full TREC document collection is 2 Gbytes of text from newspapers, newswires, patent applications etc., divided into independent documents, 742,611 for the full collection [Harm93].[7] For each of the 50 queries, the top 1000 ranked documents per TREC participant are returned to NIST and pooled together to yield, for each query, a candidate set of documents among which the relevant ones are likely to be found. These pools of documents are then manually assessed for relevance and the net result from this process is that the retrieval performance of groups participating in the TREC experiment can be measured. Such measurements are primarily done in terms of precision and recall, computed at standard recall points and interpolated over a set of queries. A follow on from all this is that there is now available a large corpus of text and a set of queries with known relevant documents, for experimental IR, something that was not in place for the information retrieval research community until quite recently.

In TREC-2, run during 1993, there were 23 groups participating in the so-called *ad hoc* retrieval and because of the heterogeneity of the retrieval approaches used by the different groups, it is believed by information retrieval researchers that the set of relevance assessments for TREC queries is quite good given the size of the document base. Contributing groups used approaches based on n-grams, probabilistic models, boolean queries, bayesian inference, combination of evidences, NLP and query expansion so while most of the groups used approaches based on directly matching words or word variants, there were cases where relevant documents could have been found through indirect matching between query and document terms, eg via query expansion, bayesian inference, etc. The work reported here was not done as part of any formal TREC experiment.

### 4.2 Implementation of Document Retrieval

Our approach to document retrieval, as previously stated, is to get a KB representation of both the documents and queries and to compare these representations using a semantic

[7] Due to the engineering difficulties and space requirements involved we restricted our experiments to the 154,000 articles making up the Wall Street Journal (WSJ) corpus

similarity function. In order to arrive at these KB representations the text of documents and queries were processed as follows :

(a) Syntactically tag words[8]
(b) Build up collocations
(c) Remove all non-nouns
(d) Remove nouns not occurring in KB
(e) Removal of non-content bearing nouns

As previously stated, a collocation is a multi-word phrase such as *fountan_pen* or *department_of_defense*. Approximately 45% of concepts in the noun portion of WordNet are collocations. The procedure to remove non-content bearing nouns involved the removal of nouns with any of the following characteristics :

- Having two or less characters
- Appearing in a general stop list[9]
- Having a high inter-document frequency, (above 10%).

The average length of articles in the WSJ corpus is approximately 400 terms, however, following the above pre-processing procedures we reduced this to approximately 120 terms. All that was left to arrive at a KB representation for documents, (and queries), was to locate the KB synsets corresponding to these terms. This would have been a very straightforward process if it wasn't for the presence of polysemous terms and one of the hallmarks of WordNet is its capability for very fine sense distinction. On average we found that 72% of preprocessed index terms had more than one sense in WordNet with an average of 3.1 senses per ambiguous term. For an example of this see figure 4.

[8] The ENGCG syntactic parser from the Research Unit for Computational Linguistics at the University of Helsinki [Karl89] was used for this step

[9]We used the general stop list generated by Fox, [Fox90].

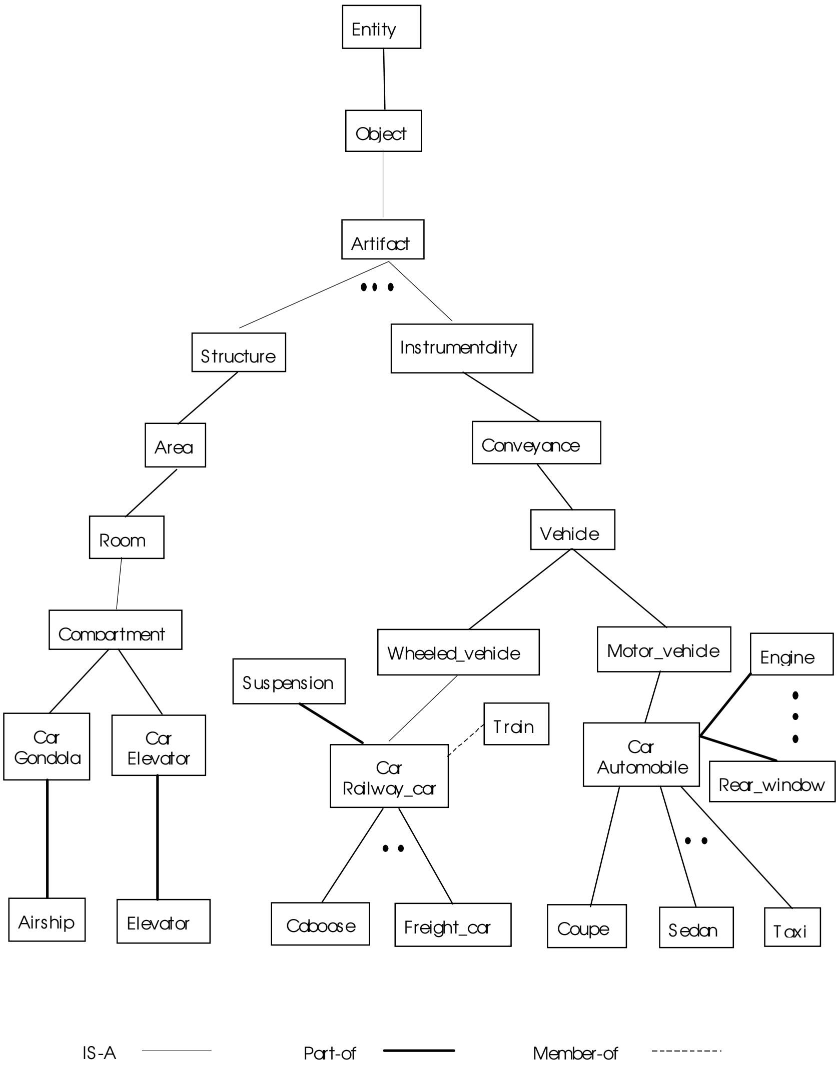

**Figure 4 KB Extract for the concept 'car' with 4 senses**

Due to the size of the collection and the extent of the ambiguity the option to manually disambiguate or indeed to simply disregard ambiguous terms was not possible. An automatic word sense disambiguator was subsequently designed and implemented. The surrounding words in the text were used as a context and we applied four different disambiguating techniques. Each technique allocated a score to each sense of the ambiguous word and the top scoring sense, (or in some cases the top two senses), were chosen as the appropriate sense [Rich94].

Having arrived at a set of KB terms representing the document and query, our first approach to measuring the similarity between them was to compute the term-term similarity between all possible combinations of document and query terms, add the similarity values together and normalise by the number of terms in the document. This approach does not take any account of clusters of terms about some concept within a query or document like the terms *engine, gearbox* and *tyres* being clustered around the concept of a car. In our approach, all terms are treated as equally important though we did weight terms from document titles slightly higher and we disregard the lowest scoring term-term similarity values. Furthermore we split large documents into equally sized units we refer to as pages. Each page is subsequently matched against the query and the similarity value of the highest scored page is used as the overall similarity score for the document to the query. This avoids unfairly penalising large documents which may discuss a number of different topics, (perhaps only one of which is relevant to the query).

A final aspect of our semantic based retrieval approach is the fact we employ a mechanism to reduce noise. Our semantic similarity estimators are, in general, quite good at estimating the similarity between terms that are relatively similar, [Rich94], however, if the two terms have not that much in common then it becomes more difficult to defend or account for the values returned. Of course, this problem could also be said to effect humans in their judgement of similarity. Most people would have little difficulty in rating the comparative similarities between a banana and an apple and between a banana and a car, however, the similarity values become more unclear if we're rating the similarity between a car and an apple and a car and a dog. All that can be said is that both are simply dissimilar. In line with this argument, we introduced absolute noise threshold values, (finalised through experimentation), to our retrieval approach. Any index/query term comparisons which are below these noise thresholds are considered to have nothing in common and are discarded from the similarity assessment of the document to the query.

In an ideal situation with infinite computing resources we would have liked to implement our retrieval strategies on all documents in the TREC corpus but because of the amount of sense disambiguation to be performed on documents, and the number of term-term comparisons to be made, each of which requires searching through our HCGs, we could not do this. Our HCGs are too large to fit into main memory using the computing resources at our disposal (a SUN SparcServer). In our experiments we used 12 of the TREC-2 queries (chosen randomly) and computed retrieval for each against a subset of 1000 documents from the Wall Street Journal segment of the TREC database. These 1000 documents for each query were composed of the top 1000 articles retrieved by a conventional IR approach based on *tf.IDF* weighting. While this may not sound like much experimentation, we had to run these queries against these documents many times when computing the optimal settings for parameters in the overall retrieval strategy. Only the most effective retrieval strategy will be presented here.

In order to give some benchmark against which our semantic distance based retrieval could be compared, we computed the performance of a more contemporary IR strategy on the document collection and queries we used.[10] In this benchmark strategy we removed stopwords and stemmed remaining words, built an inverted file and computed document scores based on *tf*IDF* weighting of keywords which automatically assigns higher weights to less frequently occurring word stems. This is a fairly standard approach to IR and is known to perform reasonably well. It is, however, based on string matching between word variants as occurring in queries and in documents, so it does offer an alternative strategy to the ones we have proposed here.

## 5. Experimental Results

The performances of our semantic based systems and that of the traditional *tf*IDF* ranking interpolated over the 12 test queries are presented in figure 4. As can be seen, the traditional *tf*IDF* ranking out-performs our semantic-based systems at all levels of recall. In a query by query breakdown of the results it was found that the conceptual distance metric performs better than *tf*IDF* in only one query and the information based outperforms the *tf*IDF* ranking in just two queries. For all other queries the performance of both semantic based systems are either comparable to or considerably worse than that of the *tf*IDF* approach. However, of particular interest from these results is the fact that the performance of both semantic based systems did not overlap. The information based system performed well for queries 1, 5, and 12, whereas the conceptual distance system performed well in queries 6, 7, and 9. This result would support a future investigation into ways of combining both approaches.

[10] Thanks are due to Fergus Kelledy for allowing us to use his work here

In the following section we put forward a number of arguments to explain our results.

| | Conceptual Dist. | Information Based | tf*IDF |
|---|---|---|---|
| Avg. Precision | 0.1062 | 0.1151 | 0.2072 |

**Table 1 Average Precision for all approaches**

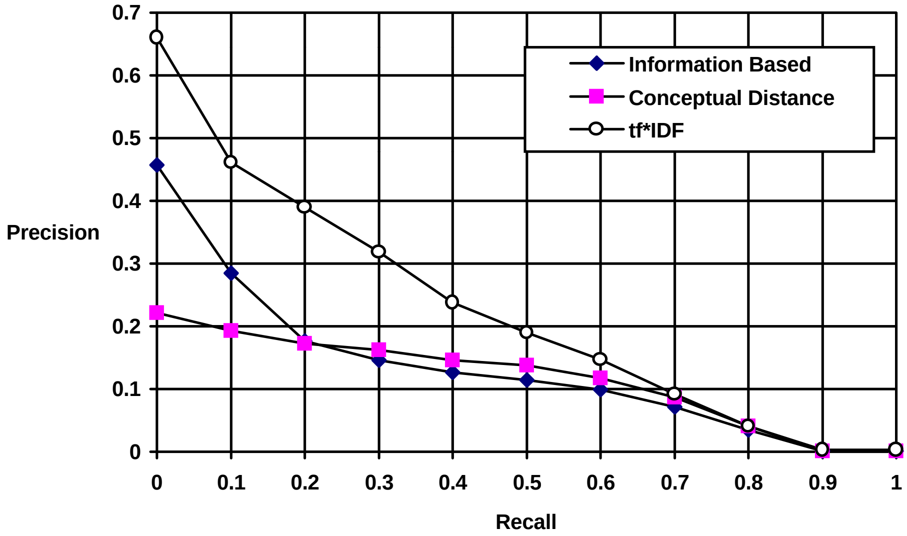

**Figure 5: tf*IDF vs. Information Based Vs. Conceptual Distance**

## 6. Analysis and Conclusions

There are many reasons which can explain the poor results of our semantic similarity based retrieval strategies and some are put forward here.

- In TREC, the mechanism used to determine the relevant documents has been criticised as not favouring approaches which do not retrieve based on word string matching. When using a retrieval strategy outside official TREC runs as we have done here, there is no guarantee that one's top-ranked documents will have been assessed for relevance though it is impossible to see how anything could be done here.
- Many aspects of our overall implementation have not been fine-tuned. For example, the word sense disambiguator has not been evaluated and we know [Krov92, Sand94] how crucial sense disambiguation is to the quality of retrieval. This compares with the performance-enhanced, (for the WSJ), variation of a standard *tf*IDF* system we used as a benchmark.
- There are many occurrences of proper nouns in TREC queries which do not occur in WordNet. In fact there is a correlation between the queries with proper nouns, and queries which performed badly for us.

- Due to constraints of time and resources our approach to evaluating our system was to use it to re-rank a set of 1000 documents retrieved by a traditional pattern matching system. One of the main strengths of our approach, the ability to relate semantically similar but lexically different terms, is not afforded the opportunity to impact performance in this experimental design. By definition the documents making the 1000 test set use query terms as index terms as opposed to using semantically related terms.

- The nature of TREC queries is that they are very detailed information need topic statements and have already proved difficult to get good results using techniques like query expansion [Voor94]. Our semantic distance based approach is similar to query expansion, trying to improve the performance of IR by bringing in or identifying related terms or words

It could be argued that using only 12 queries was not a large enough set to work with, Figure 7 shows precision-recall figures for the tf*IDF-based retrieval strategy using our 12 queries and also using the full TREc-2 set of 50 queries. From the graph it can be seen that there is very little difference between the two which supports the argument that the 12 queries we did use are fairly representative.

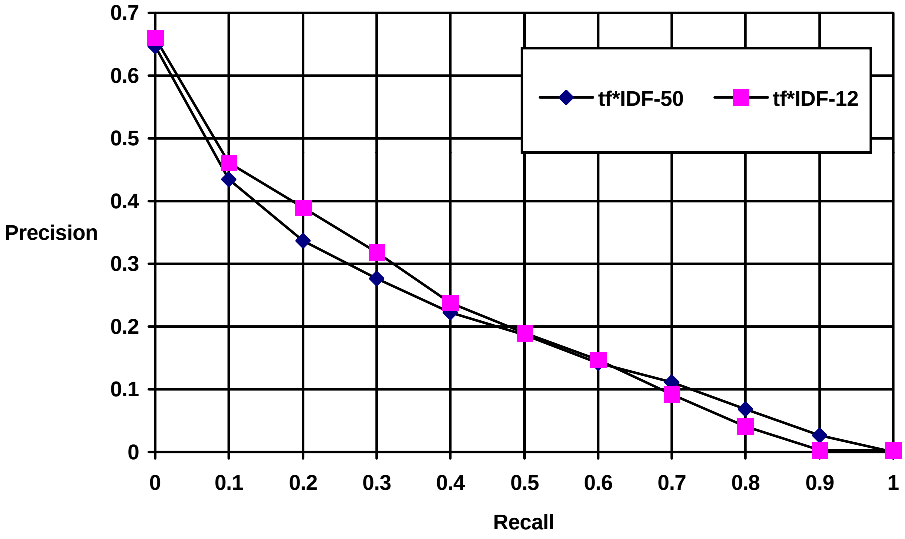

**Figure 7 Performance of the tf*IDF system over 12 and 50 queries**

In summary, our results are disappointingly not as good as conventional statistical weighting however this negative result should not be seen as wholly negative but as offering promise. Many of our queries perform very well with our strategies and notwithstanding the reasons given above, we believe our results are certainly worth pursuing.

The application of natural language processing tools and techniques to information retrieval tasks is an aspiration which is long been identified as potentially useful for the quality of IR delivered. The application of deeper levels of NLP like anahora and pronoun resolution and deep semantic analysis have not yet received much attention because of the overheads and domain dependence of current approaches and at present these seem some distance away from being used in IR. Phrase identification and NLP resources like proper name databases, name lists, thesauri etc, have been shown to be useful to IR but the area needs further work and more maturity and the issues of scale must be addressed. What we have introduced in this paper is an approach to using an NLP resource to tackle a fundamental and inherent difficulty with information retrieval, the use of synonyms and related terms in describing the content of either queries or documents. Our initial results presented here seem to have raised more questions than we have answered and we shall be pursuing this approach in future work.

## References

[Bate86] : M. Bates, “Subject Access in Online Catalogs: A Design Model”, Journal of the American Society for Information Science, 11, 357 - 376, 1986.

[Beck92] : Richard Beckwith and George A. Miller, "Implementing a Lexical Network", *Report No. 43*, Princeton University, April 8, 1992.

[Beck93] : Richard Beckwith, George A. Miller, and Randee Tengi, "Design and Implementation of the WordNet Lexical Database and Searching Software", Working Paper, Princeton University, 1993.

[Bota92] : A. Botafogo, E. Rivlin, and B. Shneiderman, "Structural Analysis of Hypertexts: Identifying Hierarchies and useful Metrics", *ACM Transactions on Information Systems*, 10 (2), 142-180, 1992.

[Furn87] : G. W. Furnas, T. K. Landauer, L. M. Gomez,, and S. T. Dumais, "The Vocabulary Problem in Human-System Communication", Communications of the ACM, Vol. 30, No. 11, November 1987, 964-971.

[Gins93] : A. Ginsberg, "A Unified Approach to Automatic Indexing and Information Retrieval", IEEE Expert, special issue on AI, 8, (5), 46-56, 1993.

[Harm93] : D. Harman, “An Overview of the First Text Retrieval Conference (TREC 1), *Proceedings of TREC 2*, 1-21, 1993.

[Kim90] : Young Whan Kim and Jinh H. Kim, "A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph", *Journal of Documentation*, 46 (2), 113-137, 1990.

[Krov92] : R. Krovetz and W. B. Croft, “Lexical Ambiguity in Information Retrieval”, *ACM Transactions on Information Systems*, 10 (2), 115-141, 1992.

[Lee93] : J. H. Lee, M. H. Kim, and Y. J. Lee, "Information Retrieval Based on Conceptual Distance in IS-A Hierarchies", *Journal of Documentation*, 49 (2), 113 -136, 1993.

[Mill90a] : George A. Miller, Richard Beckwith, Christiane Felbaum, Derek Gross, and Katherine Miller, "Introduction to WordNet : An On-line Lexical Database", International Journal of Lexicography, Vol. 3, No. 4, 1990, 235 - 244.

[Mill90b] : George A. Miller, "Nouns in WordNet : A Lexical Inheritance System", International Journal of Lexicography, Vol. 3, No. 4, 1990, 245 - 264.

[Mona87] : I. Monarch and J. Carbonell, " CoalSORT : A Knowledge-Based Interface", IEEE Expert, Spring 1987, 39 - 53.

[Rada89] : R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and Application of a Metric on Semantic Nets", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, No. 1, January/February 1989, 17-30.

[Resn93a] : P. Resnik, "Selection and Information : A Class based Approach to Lexical Relationships", PhD. dissertation at the University of Pennsylvania. Also appears as Technical Report 93-42, November 1993.

[Resn93b] : P. Resnik, “Semantic Classes and Syntactic Ambiguity", ARPA Workshop on Human Language Technology, Princeton, March, 1993.

[Rich94] : R. Richardson, “A Semantic-based Approach to Information Processing”, Ph.D. thesis, School of Computer Applications, Dublin City University, 1994.

[Sand94] : M. Sanderson, “Word Sense Disambiguation and Information Retrieval”, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, 49-57, 1994

[Suss93] : M. Sussna, "Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network", *Proceedings of the second International Conference on Information and Knowledge Base Management (CIKM),* 1993.

[Tver77], : A. Tversky, “Features of Similarity”, Psychological Review, 84, (4), 1977, 327 - 352.

[Voo94] : E. M. Voorhees, “Query Expansion Using Lexical-Semantic Relations”, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, 61-69, 1994.

## Appendix A

Calculation of Class Probabilities

Class probabilities are used in the determination of the information content or specificity of WordNet classes. The specificity of a class can be defined in terms of its class probability as follows :

$$Specificity(C_i) = -\log(P(C_i))$$

where P(Ci) is the class probability of class *i*.

In order to define the probability of a class we must first define *words(c)* and *class(w)*. *Words(c)* is defined as the set of words in all directly or indirectly subordinate classes of the class c. For example *Words(cloister)* consists of *religious residence*, *convent*, *abbey*, *friary*, *monastery*, *nunnery*, and *priory*. *Classes(w)* represents the set $\{c \mid w \in words(c)\}$ , i.e. this includes all the classes in which the word *w* is contained, regardless of the particular sense of *w*. From these two definitions we can define the frequency of a class as :

$$Freq(C_i) = \sum_{w \in words(c)} \frac{1}{|classes(w)|} \times Freq(w)$$

where Freq(w) is the frequency of occurrence of word *w* in a large text corpus. The class probabilities can be estimated from such a distribution using maximum likelihood estimation (MLE) :

$$P(c) = \frac{Freq(c)}{N}$$

where N is defined as $\sum_{c'} Freq(c')$, i.e. the total size of the sample.