

A Semantic-based Approach to Information Processing



**DUBLIN CITY
UNIVERSITY**

Ollscoil Chathair Bhaile Atha Cliath

By

Ray Richardson B. Sc.,

A Dissertation Presented in Fulfilment
of the Requirements for the Ph. D. Degree

Supervisors

Dr. Alan F. Smeaton

Dr. John Murphy

School of Computer Applications

October 1994

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. degree, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Ray Richardson Date: 23/2/95
Ray Richardson

Acknowledgements

I had often imagined writing the acknowledgements would be a simple task, but now that I'm actually doing it I realise how difficult it is to find the words to express my gratitude to all those who helped me complete what was for me a mammoth piece of work. One thing I do know is that, I am very glad to be in a position where I can write acknowledgements.

Firstly I would like to thank my supervisors, Alan Smeaton and John Murphy. I would like to thank John for his unerring faith in me. I can't remember earning this faith but I do know that at many stages throughout it was all that kept me going. I'm sure he knew failure wasn't something that would sit well with me so setting a challenge was a good tactic. I would like to thank Alan for his genuine enthusiasm in everything I did. The subject of my thesis didn't completely overlap with his research interests at the outset, however, he succeeded in managing the research very well. I am particularly indebted to him for taking the time and effort involved in keeping up to date with what I was doing. It can't have been easy as I seem to remember coming up with new strategies to various problems almost weekly. In fact it amazes me thinking back on how many times his comments, often inadvertent, would send me off in a new direction. In this way Alan provided the necessary catalyst which permits the essential element of lateral thinking in any research.

I would like to thank my friends Jim, Chris, Aiden, Ambrose, and Dec. Their support over the last few years is very much appreciated. I imagine I would probably have spontaneously combusted over a year ago had it not been for the release valve they provided.

I would also like to thank my colleagues in the postgrad lab for their support throughout. Everyone helped in different ways. I'd like to acknowledge my fellow 'IR researchers', Fergus and Ruairi and the new guys on the scene, Emmet and Ian. I hope the many conversations with Ruairi and Fergus, (ranging from the intractable problems faced by IR researchers to the bigger questions in life such as the possible side effects of consuming two canteen scones in one day), are as beneficial to them in their research as they were to me. I would also like to thank the other postgrads who, if nothing else, helped me keep a grasp on reality; John, Garry, Keith, Barry, Brian, Donal, Coleman, Willie, Pat and Kieran. I would especially like to thank John Walsh for turning on the light for me with regard to the hundreds of options in Microsoft Word.

I would also like to thank my family for their love and support throughout. I am especially grateful to them for making no demands upon me and for creating an atmosphere where I did not feel pressurised into achieving certain standards. I hope I can do the same for my children in the future.

Finally I'd like to acknowledge the significant input of Regina, my girlfriend. On many occasions I'm sure I bored her to death talking about some problem or other I was having with the thesis. But she always patiently listened to me rabbit on and more often than not helped me to see a new angle on the whole thing. I'd like to thank her for her patience, encouragement and faith in me.

Table of Contents

ABSTRACT	v
CHAPTER 1 - INTRODUCTION	1
1. Introduction	1
1.1 Motivating Application	3
1.2 Non-Semantic View of Information	6
1.3 Objective	8
CHAPTER 2 - RELATED RESEARCH	9
2. Introduction	9
2.1 Review	9
2.1.1 Contemporary Approach to IR	10
2.1.2 Linguistic Approach	13
2.1.3 Knowledge Based Approach	16
2.2 What we Propose	23
2.3 Summary	25
CHAPTER 3 - WORDNET	26
3. Introduction	26
3.1 WordNet	27
3.1.1 Semantic organisation of Nouns in WordNet	29
3.1.2 Physical organisation of information in WordNet	33
3.2 The WordNet Package	36
3.3 Current Uses of WordNet	38
3.4 WordNet the Future	40
3.4.1 Addition of Relational Pointers	41
3.4.2 Adding Context	42
3.5 Summary	43
CHAPTER 4 - KB CONSTRUCTION	45
4. Introduction	45
4.1 Building of HCGs	45

4.2 Extension to HCG synsets	50
4.2.1 Calculating a value for Information Content	51
4.2.2 Implementation Details	52
4.2.3 Improving Results	55
4.2.4 Conclusions on the Information Content value Extension	66
4.3 Weighting HCG Relational Links	67
4.3.1 Previous Weighting Strategies	68
4.3.2 New Link Weighting Mechanism	70
4.3.3 Conclusions on Weighting	73
4.4 Summary	74
 CHAPTER 5 - SIMILARITY	 76
5. Introduction	76
5.1 Related Research	77
5.1.1 Models of Synonymy	78
5.1.2 Geometric Models	80
5.1.3 Contrast Model of Similarity	84
5.2 Similarity Estimators Employed	87
5.2.1 Conceptual Distance Similarity Estimator	87
5.2.2 Information Based Similarity Estimator	91
5.2.3 Conclusions on Similarity Measures	96
5.3 Psychological Evaluation	97
5.3.1 Test data set	98
5.3.2 Initial Results	99
5.3.3 Evaluation of Weighting Strategies	100
5.3.4 Conclusions on Psychological Evaluation	105
5.4 Conclusions	105
 CHAPTER 6 - APPLICATIONS	 107
6. Introduction	107
6.1 NSD Applications	109
6.2 SD Applications	114
6.3 Choice of Evaluation Application	115
6.3.1 Availability of a test bed.	116
6.3.2 Human Factor	116
6.3.3 Evaluation Procedure	117
6.3.4 Information Volume	117
6.4 Conclusions	118
 CHAPTER 7 - GENERATION OF KB REPRESENTATIONS	 119
7. Introduction	119
7.1 Text Preprocessing	120

7.1.1 Syntactic Analysis	121
7.1.2 Processing of Syntactically Parsed text	123
7.2 Sense Disambiguation	130
7.2.1 Sense Disambiguation in Information Retrieval	131
7.2.2 Method	134
7.2.3 Sample Text	140
7.3 Possible Improvements to Sense Disambiguator	145
7.4 Summary	148
 CHAPTER 8 - RETRIEVAL ENGINE	 149
8. Introduction	149
8.1 Design of the Retrieval Engine	149
8.2 Sample Run of the Retrieval Engine	154
8.3 Conclusions on Retrieval Engine	161
8.4 Baseline Retrieval System	162
8.5 Summary	166
 CHAPTER 9 - EVALUATION	 167
9. Introduction	167
9.1 TREC	167
9.1.1 TREC Corpus	168
9.1.2 TREC Evaluation Mechanism	170
9.2 Optimum Noise Threshold	173
9.2.1 Experimental Design	174
9.2.2 Experimental Results for Noise Thresholds	177
9.2.3 Discussion of Results	181
9.3 Full Test Evaluation	185
9.3.1 Experimental Design	185
9.3.2 Results	186
9.3.3 Conclusions on evaluation	196
9.4 Summary	197
 CHAPTER 10 - CONCLUSIONS AND FUTURE DIRECTIONS	 198
10. Introduction	198
10.1 Future Directions	199
 REFERENCES	 203

APPENDICES

APPENDIX A - HCG CONSTRUCTION

APPENDIX B - CALCULATING INFORMATION CONTENT VALUES

APPENDIX C - SENSES USED IN PSYCHOLOGICAL EVALUATION

APPENDIX D - RESULTS OF PSYCHOLOGICAL EVALUATION

APPENDIX E - SYNTACTIC LABELS FOR RUCL PARSER

APPENDIX F - STATISTICS ON WSJ CORPUS

APPENDIX G - STOP LIST USED IN PREPROCESSOR

APPENDIX H - MOST FREQUENTLY OCCURRING WSJ TERMS

APPENDIX I - QUERY BREAK DOWN OF THRESHOLDING EXPERIMENTS

APPENDIX J - QUERY BREAK DOWN OF FULL-TEST EXPERIMENTS

APPENDIX K - QUERY BREAK DOWN OF BEST QUERIES

Abstract

The research reported in this thesis is centred around the development of a semantic based approach to information processing. Traditional word-based pattern matching approaches to information processing suffer from both the richness and ambiguousness of natural language. Although retrieval performances of traditional systems can be satisfactory in many situations, it is commonly held that the traditional approach has reached the peak of its potential and any substantial improvements will be very difficult to achieve, [Smea91]. Word-based pattern matching retrieval systems are devoid of the semantic power necessary to either distinguish between different senses of homonyms or identify the similar meanings of related terms. Our proposed semantic information processing system was designed to tackle these problems among others, (we also wanted to allow phrasal as well as single word terms to describe concepts). Our prototype system is comprised of a WordNet derived domain independent knowledge base (KB) and a concept level semantic similarity estimator. The KB, which is rich in noun phrases, is used as a controlled vocabulary which effectively addresses many of the problems posed by ambiguities in natural language. Similarly both proposals for the semantic similarity estimator tackle issues regarding the richness of natural language and in particular the multitude of ways of expressing the same concept.

A semantic based document retrieval system is developed as a means of evaluating our approach. However, many other information processing applications are discussed with particular attention directed towards the application of our approach to locating and relating information in a large scale Federated Database System (FDBS). The document retrieval evaluation application operates by obtaining KB representations of both the documents and queries and using the semantic similarity estimators as the comparison mechanism in the procedure to determine the degree of relevance of a document for a query. The construction of KB

representations for documents and queries is a completely automatic procedure, and among other steps includes a sense disambiguation phase. The sense disambiguator developed for this research also represents a departure from existing approaches to sense disambiguation. In our approach four individual disambiguation mechanisms are used to individually weight different senses of ambiguous terms. This allows the possibility of there being more than one correct sense.

Our evaluation mechanism employs the Wall Street Journal text corpus and a set of TREC queries along with their relevance assessments in an overall document retrieval application. A traditional pattern matching $tf*IDF$ system is used as a baseline system in our evaluation experiments. The results indicate firstly that our WordNet derived KB is capable of being used as a controlled vocabulary and secondly that our approaches to estimating semantic similarity operate well at their intended concept level. However, it is more difficult to arrive at conclusive interpretations of the results with regard to the application of our semantic based systems to the complex task of document retrieval. A more complete evaluation is left as a topic for future research.

LIST OF FIGURES

Figure 1-1 Increase in Internet Traffic for Gopher and WWW (Dec. '92 - June '94)	2
Figure 3-1 Extract from the WordNet Lexical Inheritance System	31
Figure 3-2 Senses of horse in WordNet	35
Figure 4-1 Indexing of concepts in KB	50
Figure 4-2 Single Vs Collocation concepts	59
Figure 4-3 Synsets with Information content Values Vs. those without	60
Figure 4-4 Percentage of collocations found in WSJ extract	60
Figure 4-5 Inclusion of Collocations in Information Content Calculation	64
Figure 4-6 KB Extract	70
Figure 4-7 Local Densities in HCGs	72
Figure 5-1 An example of similarity in a geometric space. (Taken from [Rips73])	81
Figure 5-2 Multiple Inheritance	89
Figure 5-3 Weighted KB Extract	90
Figure 5-4 KB Extract for the concepts 'car', 'fork', 'bicycle', and 'banana'	92
Figure 5-5 KB Extract showing violation of minimality in information based similarity estimator	95
Figure 6-1 A Sample Session with a prototype front-end	112
Figure 6-2 A sample dialogue for expanding terms	113
Figure 7-1 Average Lengths of WSJ articles	129
Figure 7-2 WordNet Extract for the concept 'car'	133
Figure 7-3 Entries for 'door'	135
Figure 8-1 Non-metric Information based Similarity Estimator	157
Figure 8-2 Weighted KB Extract	160
Figure 8-3 Structure Generated for Queries	164
Figure 9-1 Sample Precision/Recall graph Comparing two Systems	172
Figure 9-2 Optimum Noise Threshold For Conceptual Distance Configuration	179
Figure 9-3 Optimum Noise Threshold for the Information Based Configuration	180
Figure 9-4 Information Based Vs. Conceptual Distance System	183
Figure 9-5 $tf*IDF$ Vs. Information Based Vs. Conceptual Distance	187
Figure 9-6 Optimised $tf*IDF$ system Vs. Non-optimal $tf*IDF$ system	190
Figure 9-7 Performance of the $tf*IDF$ system over 12 and 50 queries	193
Figure 9-8 Performance of all three systems on 4 best thresholding queries	195

LIST OF TABLES

<i>Table 3-1 Encoded Relational Pointers</i>	34
<i>Table 3-2 Hypernyms of bronco and their familiarity values, taken from [Beck93].</i>	35
<i>Table 3-3 Inflectional endings for Nouns</i>	37
<i>Table 4-1 Distribution of collocations in HCGs</i>	59
<i>Table 4-2 Unique collocations detected in different sized WSJ extracts</i>	62
<i>Table 5-1 Average Human Similarity Scores for 30 Noun Pairs</i>	98
<i>Table 7-1 Test Articles</i>	140
<i>Table 8-1 Term Weightings</i>	151
<i>Table 9-1 Source material for TREC document Collection</i>	169
<i>Table 9-2 Overlap of Submitted Results</i>	170
<i>Table 9-3 Ranked List of 20 retrieved documents</i>	171
<i>Table 9-4 Sample Precision and Recall Values</i>	171
<i>Table 9-5 Initial Absolute Noise Threshold Values</i>	177
<i>Table 9-6 Average Precisions for each configuration</i>	178
<i>Table 9-7 Average Precision for all systems in the full test</i>	186
<i>Table 9-8 Average Precision for all systems on 4 best thresholding queries</i>	194

GLOSSARY OF ACRONYMS

BT	BROADER TERM
FDBS	FEDERATED DATABASE SYSTEM
HCG	HIERARCHICAL CONCEPT GRAPH
IDF	INVERSE DOCUMENT FREQUENCY
IR	INFORMATION RETRIEVAL
KB	KNOWLEDGE BASE
KBIR	KNOWLEDGE BASED INFORMATION RETRIEVAL
MLE	MAXIMUM LIKELIHOOD ESTIMATION
NLP	NATURAL LANGUAGE PROCESSING
NSD	NON SELF-DESCRIBING
NT	NARROWER TERM
RT	RELATED TERM
SD	SELF DESCRIBING
SDM	SEMANTIC DISTANCE MEASURE
SSM	SUMMARY SCHEMAS MODEL
TF	TERM FREQUENCY
TREC	TEXT RETRIEVAL CONFERENCE
WSJ	WALL STREET JOURNAL
WWW	WORLD WIDE WEB

Chapter 1 - Introduction

1. Introduction

The world is currently experiencing an information explosion. The proliferation of cheap storage devices and computing power has meant there is more information online than ever before and the rate at which information is being generated is growing at an enormous rate. Coupling these developments with recent advances in networking and telecommunication technologies, the amount of information accessible to individuals and organisations has reached almost incomprehensible proportions. New services such as the World Wide Web, (WWW), a globally-distributed freely available hypermedia system, allow users to travel seamlessly through the Internet cyberspace. Such services have also extended the traditional textual view of information to include multimedia information such as sound, images and video. According to recent estimates, the Internet is gaining roughly 150,000 new users per month, joining the 20 million existing Internet users, [Pitk94]. This increase in users is reflected in an increase in the amount of traffic on the Internet, see Fig. 1.1.

However, a single looming problem is becoming apparent with the advent of this new information age. This problem is concerned with adequate resources to process this information. A staggering 1,000,000 person years per year is spent, (just in the European Union alone), searching for information, [Smea94]. This is just one facet of information processing, and it is not difficult to see how heavy the demands processing information are becoming for individuals and organisations. The modern world is critically dependent on information, and given the vast quantities of it, critically dependent on computers and computing techniques for handling it. However, the current computational approaches to processing information are overly simplistic and consequently failing in their task. A new approach is urgently required.

We readily acknowledge, however, the immensity of work involved in such an endeavour. In our research we address only some of the problems encountered by current approaches to automatic information processing.

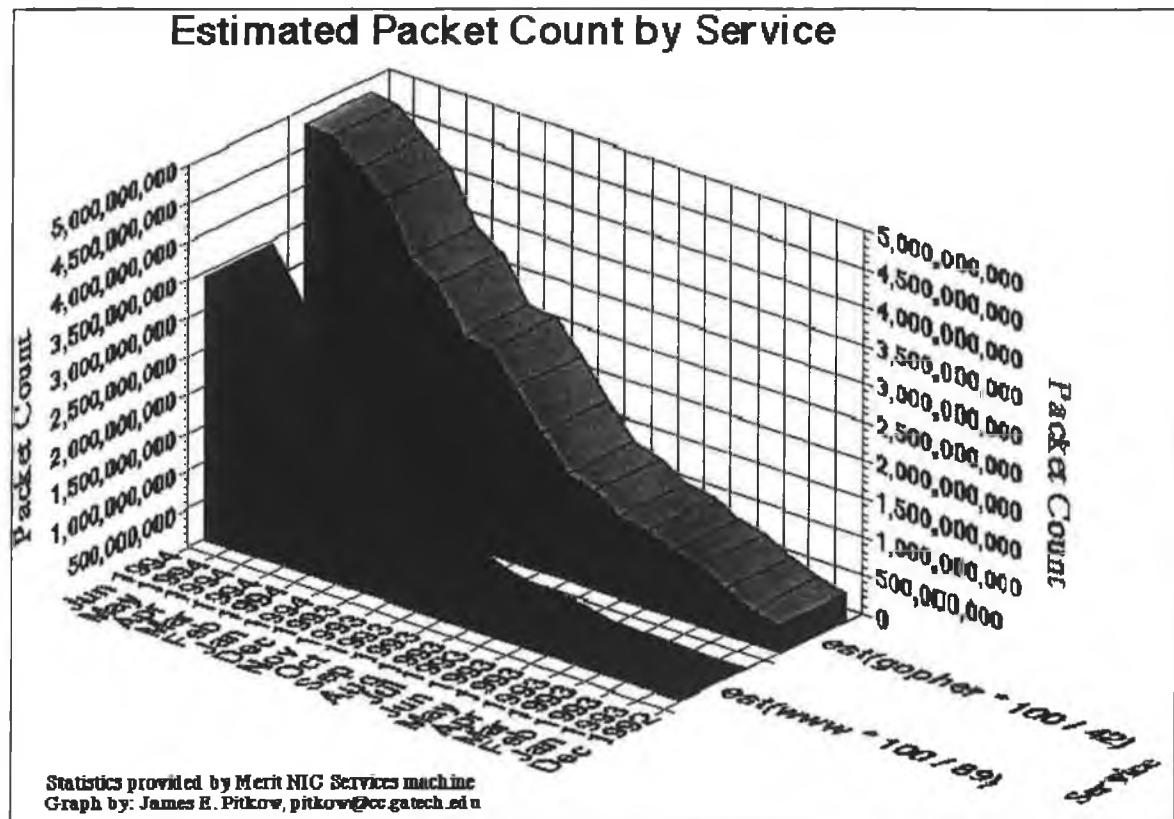


Figure 1-1 Increase in Internet Traffic for Gopher and WWW (Dec. '92 - June '94)

In the following Section we describe in a little more detail some of the problems arising from the current information explosion within a specific domain. In Section 1.2 we briefly present some of the short comings of the current approaches to automatic information processing. Finally in Section 1.3 we outline the objectives of this research.

1.1 Motivating Application

One category of information not alluded to in the previous Section is the structured information found in computer databases. Prior to now, dealing with this information was quite straight forward given its *structured* form, however, this kind of information has also witnessed its own information explosion. Traditionally, individual organisations maintained their own centralised databases which created an environment conducive to standardisation in terms of the choice of database model, query language and naming policy for database schemata, (a set of access terms used to describe the structure and content of data available in a database). However, corporate decentralisation coupled with networking and telecommunication advances during the eighties changed the focus from centralised databases to decentralised or distributed databases. With decentralisation came autonomy of local databases both in terms of choice of database models and naming policies for schemata. These developments have led to many problems in current business activities. Incompatibility and lack of standardisation has become a major problem in the event of company take overs, company mergers, and in changes of company policy where by global observing bodies are introduced to either act as a governing body or more importantly to promote sharing of information across decentralised sites.

As such, there is an express need to share data stored often in heterogeneous database systems. However, contrary to the centralised database approach, there is also a desire to preserve component database autonomy. Users argued that total integration was not always necessary or even desirable. In situations where an organisation has significant investment in hardware, software, and user training, it is understandable that the organisation would wish to preserve the investment by ensuring existing local applications continued to operate unchanged. Organisations often want to preserve the autonomy of each database, even to the point of refusing to participate in a globally integrated schema. This is particularly true in the sharing of information between organisations, or in public information services like Teletel in France, [Cats94], where the owners of component databases compete for customers.

To satisfy both requirements, current research efforts are focusing on the Federated database (FDBS) approach.

The term federated database was first introduced in [Heim85] and is described in [Shet90] as a collection of co-operating but autonomous distributed database systems. There is no centralised control since component databases control access to their own data. This control extends to deciding which data is to be made globally available. Component database administrators produce what is known as an export schema which describes what data in their database is to be made available to other databases in the federation. Federated schema(s) are then built by combining the export schemas. The method of building the federated schema(s), i.e. through schema integration or importation, coupled with the knowledge of who has responsibility for their creation and maintenance determines the category of the resulting FDBS. Simply stated, in the tightly coupled approach a central authority, commonly referred to as the federated database administrator, is responsible for creating federated schemas and the process itself takes the form of schema integration. In the loosely coupled approach each user is the administrator of his/her own federated schema and the process of developing federated schemas is through schema importation.

Research into heterogeneous databases has traditionally investigated approaches to sharing data among a small number of component databases. In such an environment the integration of all export schemas to form one or more federated schemas, as in the tightly coupled FDBS approach, seems quite reasonable. Similarly the issue of "knowing what is out there" for users of loosely coupled systems is not significant since it can be assumed that the amount of information the user has to know is kept at a reasonable level. However, as the number of databases grows the environment becomes less amenable to both approaches. In the case of the tightly coupled approach the complexity involved in integrating thousands of export schemas would be enormous not to mention the practically intractable problem of maintaining the integrity of the federation given the freedom of component databases to update their local data at will. It is similarly unreasonable to assume that a user of a loosely

coupled FDBS knows exactly what he/she is looking for and what each database contains in an environment that consists of thousands of databases.

As such, neither of the current approaches to interoperability within heterogeneously distributed databases scale up adequately. However, whereas the problems of the tightly coupled approach seem intractable, it seems at least possible to work towards solutions to the problems of the loosely coupled approach. The problems of the loosely coupled approaches are centred on their added complexity for users and the requirements that users have a knowledge of data locations and relations between data sets. There is a consensus in [Silb91], [Boug92a] and [Brig92] that the problems of locating and relating information within a large number of interconnected autonomous heterogeneous databases are major open research issues. The problem extends to finding the range of subjects about which information is stored in the federation. Most commercially available systems [Brig92], provide users with the ability to scan the available data sequentially. This is, however, impractical for large systems. [Boug92b], in a separate analysis of the current state of federated databases, describes how locating information is achieved in two steps. First, the requesting database consults the federation dictionary for existing databases and available schemas, and second, imports all known schemas (whenever possible) and browses through them for a certain information type. This is quite obviously inadequate if there are tens (not even hundreds or thousands) of such schemas.

It would appear we need a system which can identify the location of specific data when given imprecise identifiers, (i.e. identifiers which don't refer to the names of structures within which data are stored). In light of the problems raised due to size it seems impractical that the item of data sharing should be import schema objects¹. Given the possibly huge amounts of data globally available, users could not be expected to be aware of the names of individual import schema objects nor would it be possible to allow users to browse through all export schemas. Furthermore, given the lack of standardisation in the naming of schema objects, it is not unusual to find

¹ If the import schema was in the relational model then an import schema object could be a global relation or attribute

objects such as 'Table0051', 'cde090', etc. These names do not in anyway reflect the nature of the information contained within the data structures.

The original motivating application for this research was thus to tackle the problem of locating and relating information in large scale FDBSs. We proposed a concept level approach to the problem, that is to allow users refer to information of interest in terms familiar to them. In the formulation and implementation of our approach we endeavoured to develop a generic information processing system, independent of any specific application. In retrospect we found that evaluating our system in an FDBS application was very difficult so we opted to evaluate it on a different application. The number of applications to which a concept level approach could be applied is enormous (e.g. document retrieval, information filtering, multimedia information retrieval as in caption retrieval, generation of hypertext tours, etc.). As shown in Chapter 6, we concentrated on document retrieval, however we will return to the motivating application, FDBS, later on.

1.2 Non-Semantic View of Information

Given that we have defined our problem and we have outlined our approach to the solution, an obvious question is what is preventing current approaches to handling information from adopting our approach ? The answer to this question can be found in the complexity of natural language. Humans can perform intelligent processing of information, however, we are incapable of handling large amounts of information. In contrast, computers can deal with vast quantities of information once there is only very limited intelligent required. Reasons can be found in the complex representation of information. Almost all information can be represented in natural language, this richness of natural language, however, makes it very difficult to process computationally. As we shall see in Section 2.1, any large scale, domain independent computational processing of information has traditionally involved a pattern

matching process, a literal character by character comparison of the word level components of natural language texts representing information. This simplistic, non-semantic view of information² effectively rules out any intelligent automatic processing. On a very broad level one can pinpoint two areas not addressed in the non-semantic approach to information processing :

(a) Natural Language Ambiguity.

(b) Richness of Natural Language.

Firstly, natural language is notoriously ambiguous, even ignoring complex expressions in natural language, where syntactic and semantic ambiguity can be involved, the very existence of homographs introduces ambiguity at the very simplest level of the single word. Homographs are words with the same spelling but different meanings, (e.g. *bank*, is it a commercial *bank* or a river *bank* ?). Humans deal very well with this form of ambiguity by inferring the correct meaning from the context. However, this inference process assumes the existence of a common sense knowledge base, something sadly lacking in computers. The computational approach of direct pattern matching is entirely too simple to deal with even the simplest forms of natural language ambiguity.

The second point regarding the richness of natural language, is concerned with the multitude of different ways to describe the same thing. Bates, in [Bate86], points out :

‘..the probability of two persons using the same term in describing the same thing is less than 20%’

and in a subsequent study, Furnas et al. write :

² Word-level processing of information is not necessarily completely devoid of semantics. It can perhaps be more correctly thought of as being a low level semantic approach to information processing

‘..the probability of two subjects picking the same term for a given entity ranged from 7% to 18%’, [Furn87].

From these findings, attempting to directly match natural language terms or words against each other in any information processing task would be expected to give bad results.

Our approach to automatic information processing can thus be further refined by stating the fact that a semantic approach will be adopted in place of a pattern matching, low level semantic approach.

1.3 Objective

The problem being tackled in this research is the domain independent, semantic processing of information on a large scale. We do not restrict information objects to be textual but rather that they can be described by natural language. This allows us to deal with structured information as in database tables and multimedia information such as sound or images. Information objects such as these can be described by captions of natural language text or through interactive dialogue with a user. The approach proposed in our research is to replace the current approaches to the computational handling of information by a procedure which can distinguish different senses of terms and can relate concepts that are semantically similar.

Chapter 2 - Related Research

2. Introduction

The majority of the research discussed in this Chapter falls into the field of Information Retrieval (IR). Certainly IR is a well established field of research and in many respects its aims are very close to the objectives of our research. The subject of IR involves the development of computer systems for the storage and retrieval of textual information, [Salt89]. This differs slightly from our aims insofar as we do not limit information to be textual nor do we see retrieval as the only operation that can be automatically carried out on information. However, all information can be represented in natural language and in this sense IR is very close to what we are doing. Also, information retrieval and what it involves accounts for the large proportion of processing carried out on information.

In the following Section we give an overview of current work being carried out in IR. In Section 2.3 we present our approach to information processing and relate it to what was presented in Section 2.2. Finally, in Section 2.4 conclusions and an overview of the layout of the remainder of the thesis are presented.

2.1 Review

IR is a very large field of research and our review is merely representative of those IR systems deemed relevant to our research. For more comprehensive overviews of current work in IR refer to [Smea92a, Crof93, Will94]. Even within our

restricted area of interest, it is still quite difficult to arrive at a taxonomy that embraces all pertinent IR systems. However a possible taxonomy would include the following approaches :

- Contemporary
- Linguistic
- Knowledge Based

In the following subsections we will discuss each of the above approaches to IR in more detail. For each approach the discussion will concentrate on issues such as the generality of their domain of applicability, their usefulness in non-textual information processing, their handling of ambiguity and problems posed by the richness of natural language, and whatever individual weaknesses that might be inherent in their approaches.

2.1.1 Contemporary Approach to IR

Almost all commercially available IR systems fall into this category, the most well known of which includes STATUS/IQ, Personal Librarian, and SMART, [Buck85]. In this approach to information retrieval textual documents are represented by a select set of words from the document, (referred to as index terms), and the relevance in a document to a query is determined by direct pairwise comparison of words of a query against index terms. One of the basic premises behind the traditional approach is the supposition that the content of a document can be captured to some extent by the frequency of occurrence of words in the document. Words which occur very frequently have a poor discriminatory value and, as such, are not good at highlighting the information content of a document, [VanR79]. Similarly, words that occur very infrequently tend not to be used as query terms. Hence, terms with midranging frequencies within the collection of documents are used to represent the document. In some systems, index terms are literally selected in this way while in

others the process is taken a step further and the frequency of occurrence of terms within a document, (the term frequency), is also taken into account. Typically, the term frequency is used to weight index terms with regard to their importance in describing the document.

As was mentioned above, the comparison mechanism in traditional IR systems is direct pattern matching between terms. Index and query terms are usually stemmed to reduce different variations of the same word to a common base form. This ensures that an index term such as *cars* finds a match with a query term such as *car*. Of course, word stemming will not allow for a situation where the index term is *automobile* and the query term is *car*.

In contemporary IR queries can take the form of a Boolean expression, a natural language statement, or simply a list of index terms. In the case of Boolean expressions the logical operators AND, OR, and NOT are used to connect query terms. Retrieval systems based on the Boolean model permit very precise queries, however trained searchers are generally required to construct queries. Also such statements cannot rank retrieved documents in terms of their perceived relevance. Documents are simply relevant or not relevant. In situations where hundreds of documents may be deemed relevant to a query, users have to sift through all retrieved documents to locate the most relevant ones or reformulate their query to make it more precise. To counter this problem, retrieval systems referred to as 'best match', [Will94], systems were developed. In this approach a query takes the form of a natural language statement or simply a list of weighted or unweighted concepts. Similarity measures such as the dot product and normalised variants, (e.g. the cosine, (1) below, and dice similarity coefficients, [Salt83]), are typically used with binary or unweighted query terms.

$$\text{Cosine}(Term_i, QTerm_j) = \frac{\sum_{k=1}^n (Term_{ik} \cdot QTerm_{jk})}{\sqrt{\sum_{k=1}^n Term_{ik}^2 \cdot \sum_{k=1}^n QTerm_{jk}^2}} \quad (1)$$

where

$Term_i$ is the i^{th} index term

$Qterm_j$ is the j^{th} query term

n is the total number of index terms representing documents

In the case of weighted index and query terms the similarity measure often used is referred to as the $tf*IDF$ (term frequency/inverse document frequency) measure. As we have seen above, the term frequency is used to weight index terms, and similarly the inverse document frequency, (IDF):

$$IDF(Q_i) = \text{Log}\left(\frac{N}{F(Q_i)}\right) \quad (2)$$

where

$Q_i = i^{th}$ query term,

N = Number of documents in collection, and

$F(Q_i)$ = the number of documents the term Q_i occurs in.

can be used weight query terms. The rationale behind IDF weighting is that people tend to use broadly defined, frequently occurring terms when defining their information needs so any more specific, i.e. low frequency terms, are likely to be important. $tf*IDF$ systems operate by weighting all index/query term matches by the $tf \times IDF$ value and then adding all these weights together to arrive at an overall score for the relevance of the document to the query. Both the weighted and unweighted similarity measures allow ranking of output in terms of their importance to the query.

In general, current IR systems can operate effectively in any domain. The mechanism of using term occurrences to index documents and term frequencies to possibly weight index terms ignores any semantic meaning of terms and thus, once there is sufficient text to describe documents, the procedure will operate successfully regardless of the document domain. This fact readily explains the popularity of the

pattern matching approach to IR. However, it is commonly held that the traditional approach has reached the peak of its potential and although small improvements in performance are still possible, any substantial improvements will be very difficult to achieve, [Smea91]. The main problem with the approach has to do with its word based non-semantic view of information. The dual problems of ambiguity and the richness of natural language place a limitation on the achievable performances of current IR systems, (refer to Section 1.2 for a further discussion on the non-semantic view of information). Also, an inherent assumption in the current approach to IR, which indexes documents by single terms, is that all concepts can be represented by a single word. Clearly this is not the case. Single concepts such as *Information Retrieval* and *Object Oriented Database* are just two examples. Research aimed at addressing these problems accounts for a large proportion of IR research currently being carried out. Word sense disambiguation techniques are being developed to tackle the problem of ambiguity, [Krov92], and thesauri are being used to expand queries to include terms and phrases related to the original query terms. [Voor94]. Thus far, however, the results of these approaches have proven to be very disappointing, [Sand94]. Indexing texts by phrases based on their statistical properties, [Crof91], has similarly not yet been shown to be productive.

A final weakness of the traditional approach, relating directly to our research relates to its application to the processing of non-textual information objects. The frequencies of words obviously differs between document collections and if the task involves retrieving captioned or user described information objects, (such as database tables or a bitmap image), quite clearly there is no way of applying the commonly used statistical techniques of the traditional approach to IR.

2.1.2 Linguistic Approach

The linguistic approach to information retrieval, [Smea92a] involves applying the techniques of Natural language processing, (NLP), to the problem of locating

information. One basic idea is to carry out a syntactic analysis of the text of both the query and the body of information to be searched, for noun phrases, and to use these phrases in the matching process. This addresses some of the problems posed by a word based approach to information retrieval as employed by the statistical approach to IR. Specifically it addresses problems posed by the assumption that all concepts can be described by a single term, (refer to the previous Section).

The syntactic analysis of text typically involves three phases :

- Morphological : Reducing words to their base forms, for example prefix and suffix removal.
- Lexical : The determination of the part of speech of a word.
- Syntactic : The determination of a word's role in a sentence, for instance what clause is it part of and is it a head or a modifier, etc..

There is scope for ambiguity at every phase of syntactic processing. In the morphological analysis of the word *axis* it is not clear whether to return the base word *axe* or the word *axis*. Lexical ambiguity can be seen in such sentences as '*I saw her duck*' - is *duck* a noun or a verb? Syntactic ambiguity can have many forms. The following examples³ illustrate just some of these forms :

- *Computer and telephone network* - is this a computer network and a telephone network or is it a computer and a telephone network.
- *I saw the boy with the telescope* - was the telescope used to see the boy or was the boy holding a telescope.
- *Computer performance evaluation* - is this the evaluation of the performance of a computer or is it performance evaluation using a computer.

³ The first example is referred to as co-ordination in a compound noun phrase, the second is called propositional phrase attachment, and the third is known as a compound noun phrase where the ambiguity is with respect to the head clause.

Morphological and lexical ambiguities are addressed by a part of speech tagger. Some taggers attempt to disambiguate by using the surrounding words [Bril93], and others simply encode all ambiguities [Karl89]. Different IR researchers have adopted different approaches to dealing with syntactic ambiguities. Some have chosen to ignore the ambiguities and to select the “most likely” alternative when ambiguity is encountered, for example [Salt90]. Others have attempted to normalise the ambiguities by matching ambiguous phrases against manually constructed phrase lists [Evan91], and others have opted to encode the ambiguities and to use them as a weighting mechanism in the query comparison process, [Sher92 and ODon94].

The comparison mechanism in the linguistic approach is usually pattern matching between word base forms. However, as well as matching individual words against each other, linguistic based retrieval systems can match phrases. Phrases can be represented by tree-like syntactic structures, and the comparison mechanism can entail a tree matching exercise in which inexact matches are permitted. An inexact match can be made if two trees differ with respect to the number or position of terms or we may have semantically similar phrases but different tree representations due to different interpretations of syntactic ambiguity.

As with the traditional approach to IR, the syntactic based linguistic approach is domain independent. A lexicon and a grammar are all that are required to syntactically analyse text. In contrast, a comprehensive knowledge base is required for a semantic analysis, [Smea91]. Unlike traditional approaches to IR, the linguistic approach could possibly be used in the processing of information described by captions or through dialog with the user. Descriptions of this kind use natural language and, as such, are open to analysis by a linguistic IR system. Retrieval systems adopting the linguistic techniques are, in general, still just research prototypes and a complete evaluation of the approach has still to be carried out. However, it does seem apparent the linguistic approach will suffer badly from the presence of synonymous phrases in natural language. This is another aspect of the richness of natural language, i.e. the multitude of ways of expressing the same idea. The phrases *stomach pain* and *post-prandial abdominal discomfort* both express the same

concept⁴, however, since they share no words in common, equivalence can only be determined with the aid of phrasal thesauri. Phrasal synonyms of this kind are very commonplace in natural language, presenting a far larger problem to linguistic approaches to IR than the related problem posed by synonyms in traditional IR systems. Synonyms can exist at the phrase and sub-phrase levels as well as at the word level. For example :

large automobile
large car
big automobile
big car

A phrasal thesaurus is needed to overcome the difficulties presented by synonymous phrases, however, 'constructing a phrase thesaurus is a huge task' [Smea91].

2.1.3 Knowledge Based Approach

Knowledge based information retrieval, (KBIR), is a new and promising approach to IR aimed at overcoming the problems of a non-semantic view of information. The basic approach in all instances is to use a controlled vocabulary to represent documents and then to represent information needs in this controlled vocabulary, [Paic91]. Traditionally, KBIR systems have been very domain specific. The MEDLINE online medical information system, [McCa80], is one of the first examples of a KBIR system. The knowledge base in MEDLINE, MeSH (Medical Subject Headings), is made up of approximately 15,000 indexing terms arranged in a hierarchical structure of nine levels of depth. The relational links between nodes in MeSH are described as *Broader Term/Narrower Term* relations. In the original manual approach, a trained indexer scanned each article and chose a set of index terms from MeSH that together described the contents of the article. The querying process involved trained searchers who accepted natural language queries and produced equivalent Boolean queries made up of MeSH terms. Finally the retrieval of documents in response to a query was accomplished by a process of direct pattern

⁴ Example taken from [Evan91].

matching of query terms and index terms in a boolean combination. However, much research has since been conducted on trying to automate the indexing and querying phases and to improve the retrieval phase. Rada developed a system called INDEXER, [Rada88], which attempted to extract MeSH index terms from article titles automatically. Given the narrowness of the domain, word sense disambiguation did not present a major problem and INDEXER was found to simulate accurately the performance of manual indexers. In [Rada89], a conceptual distance metric is proposed to replace pattern matching as the comparison process. Although pattern matching in KBIR doesn't suffer from the problems posed by homographs it does fall victim to documents indexed by terms related to query terms. Rada's approach was to use the sum of edge weights along the shortest path between two nodes as a measure of the conceptual distance between these nodes. This metric was then used to compare the equivalence of a set of query and document terms.

In a separate project, Pollitt developed CANSEARCH, an expert system to aid in the generation of legitimate query terms for querying MEDLINE⁵ on articles relating to clinical cancer therapy. A hierarchy of medical subject headings is presented to the user in the form of a menu system and the user performs the querying by selecting menu options, via a touch terminal. The domain knowledge used in CANSEARCH consists of general knowledge on clinical cancer therapy, knowledge of the MeSH controlled vocabulary of terms used to index cancer therapy documents, and knowledge of specific indexing instructions. This knowledge is visible in the workings of the user interface or hidden in the rule base which controls the functioning of the system. The system operates by matching the antecedents of rules against either user selections or internal messages on a system blackboard, and performing actions appropriate to the situation signified by the match. The eventual outcome of the process is a set of MeSH index terms describing the initial user information need. The system does not perform actual searches.

⁵ Legitimate is used to refer to terms from the controlled vocabulary used to index documents in the MEDLINE database.

Another domain specific KBIR system, CoalSORT [Mona87], was designed to aid in document indexing and query formulation in a bibliographic database at the Pittsburgh Energy Technology Centre. The application domain was restricted to catalyst applications in coal liquefaction. A frame-based semantic network, representing an expert's domain knowledge, characterises the system intelligence. Relationships between frames in the network are defined by the slots in the frames. Examples of slots in a network frame include *sub-category-of*, *sub-categories*, *examples*, *also-called*, *description*, and *parts*. The system attempts to use this network to aid indexers in the choice of index terms to catalogue documents. The same knowledge is also employed by users in the selection of query terms, thus fulfilling the requirement that queries and documents are represented by the same vocabularies. The system communicates the knowledge base to both indexers and searchers via a graphically oriented user interface. Basically the slot of the root frame of the network, called the coal liquefaction concept, is initially displayed to the user. The user is allowed to select a highlighted slot values and the system subsequently displays the associated frame. The user thus navigates his/her way through the network, selecting various concepts for search or query terms, and assigns them appropriate weights.

In both systems described above, CoalSORT and MEDLINE, the domain is well bounded and specific knowledge with respect to that domain is used to raise the level of performance of the system. However, the larger problem of applying KBIR techniques to domain independent applications is still an area of intensive research. Shoval was one of the first researchers to suggest the application of knowledge bases to broader domains, [Shov85]. Shoval's system is quite similar to CANSEARCH in its aims. It was designed to accept a user's information need and suggest a set of appropriate terms to represent his or her problem. At the heart of Shoval's system is a densely linked semantic network which is operated upon by the set of rules: Expand, Match, Suggest and Backtrack. The design of Shoval's system was quite elaborate, in terms of spreading activation rules and ranking metrics. However, the existence of a richly connected semantic network was simply assumed. As has been discovered by many researchers in KBIR, the development of such a knowledge base has proven to

be a significant obstacle to the application of KBIR in broader domains. In [Chen92] an attempt is made to automatically construct a knowledge base to be used as an aid in the retrieval of information on the general subject of East-Bloc computing. The knowledge base was constructed from a statistical analysis of 200 MBytes of manually indexed textual information on East-Bloc computing. This textual database was compiled by the Mosaic research group at the University of Arizona. The analysis amounted to a SMART-like selection of index terms from each document. All terms which appeared at least three times in the database were included as knowledge base concepts. For each such concept its term co-occurrence probabilities with all other concepts was computed. Two separate algorithms, referred to as the symmetric and asymmetric algorithms, were used to compute the co-occurrence values.

(a) Symmetric Algorithm

$$Co_occ_weight(T_j, T_k) = \frac{\sum_{i=1}^n d_{ij} \times d_{ik}}{\sqrt{\sum_{i=1}^n d_{ij}^2 \times \sum_{i=1}^n d_{ik}^2}}$$

(b) Asymmetric Algorithm

$$Co_occ_weight(T_j, T_k) = \frac{\sum_{i=1}^n d_{ij} \times d_{ik}}{\sum_{i=1}^n d_{ij}}$$

$$Co_occ_weight(T_k, T_j) = \frac{\sum_{i=1}^n d_{ij} \times d_{ik}}{\sum_{i=1}^n d_{ik}}$$

Where T_i represents concept i ; T_k represents concept k ; n represents the number of documents in the database; d_{ij} represents concept T_j in document i (value: 0 or 1); d_{ik} represents concept T_k in document i (value: 0 or 1). The end result was two separate knowledge bases.

Performance of both knowledge bases was evaluated before any integration with the information system took place. Evaluation was with respect to concept recall and concept precision measures and the performance of humans in the same experiments was used as a baseline. These evaluation measures are modifications of the traditional recall and precision measures used in information retrieval. Instead of measuring the recall and precision of documents retrieved, they measure the recall and precision of associated concepts generated by subjects and those generated with the aid of knowledge bases, in response to a source concept. The main result of the experiments were :

- Knowledge bases produced more terms than subjects in the recall test
- Term consistencies among subjects and between subjects and knowledge bases was low (in agreement with findings in [Bate86]).
- The knowledge base produced by the asymmetric algorithm performed better than that produced by the symmetric algorithm.

The knowledge base was subsequently integrated into the information system where it was employed as an aid to query formulation and augmentation. The knowledge base was also used to check the semantic completeness of indexes assigned to documents in the indexing phase.

Novel aspects of Chen's system could be said to be the broadness of its application domain, the algorithms used to automatically construct the knowledge base, and the strategy used to evaluate the performance of the knowledge bases. It is mentioned in [Chen92] that they applied their knowledge base construction methodology in another information domain. A collection of documents in the areas of database management systems and information retrieval was extracted from the DIALOG database and used in place of the Mosaic documents in the generation of a knowledge base. The results of the experiment were not reported beyond the fact they proved the applicability of their approach in other domains. Nevertheless, their approach is quite obviously not applicable in situations where a document collection

is not available, as in the processing of non-textual information. Also, the prospect of developing a single large domain independent knowledge base using this approach is very ambitious. However, perhaps the single largest problem with Chen's approach is the scarcity of semantic relations in the resultant knowledge base. The knowledge bases constructed from term co-occurrence statistics in Chen's approach are nowhere near as semantically rich as the one suggested by Shoval, or for that matter, those used in MEDLINE or CoalSORT.

A third system addressing issues relating to knowledge based information retrieval within a broad domain is described in [Gins93]. The system, referred to as WorldViews, uses a broad domained thesaurus to automatically index and retrieve information from electronic news articles as well as abstracts of technical reports from Bell Labs and other organisations. The thesaurus was manually constructed and consists of 3000 nodes or subject headings which are connected by broader term (BT), related term (RT), and narrower term (NT) links.

In the automatic indexing process thesaurus terms are automatically assigned to documents as content descriptors. The indexing is carried out in two phases. The first phase basically parses the document searching for explicit thesaurus term references. The correct term for an ambiguous concept is determined by calculating the distance, (within the thesaurus), between ambiguous senses and other terms in the same paragraph of the document. In a manner which is similar to Rada's approach, distance is estimated by the number of BT/NT links between concepts. If the concept does not exist in the thesaurus it is put in an inverted file which can be used to retrieve information as in the traditional keyword retrieval process.

In the second phase the system uses the list of explicit concept references produced in the first phase to generate a list of implicitly referenced concepts. A form of constrained spreading activation, starting at the explicitly referenced concepts and working up through BT/NT links, is used to generate this list. Since the thesaurus contains no cycles and all upward chains stop at the root, the process is guaranteed to stop. Each node tracks the number of times it is visited and the frequency of visits for

the root node is called the *total content*. Once the activation is complete the system indexes the document with the explicit concepts and with the concept nodes that were visited at least twice. The relevance of these nodes with respect to the document is estimated by dividing the frequency of visits to it by the *total content* and multiplying by 100.

Querying involves submitting a concept describing the information need. The thesaurus is checked for the concept. Any ambiguities is highlighted and the user is asked to disambiguate. The postings for the identified thesaurus concept may then be used as a reply to the query. More complex queries, involving a number of concepts, are handled by computing the interSection of the document postings contained for every term in the query. Query expansion is also facilitated by suggesting sub-topics of query terms, with the aid of the thesaurus, during the query process.

There is no mention in the paper of how the WorldView thesaurus was constructed. However, with just 3000 nodes and only NT, BT and RT links between nodes, its obviously not as broad or detailed in its domain coverage as would be required for our research. Also, with only the BT and NT links being used in the WorldViews distance function it may prove to be semantically deficient in the broader application domain.

In general, the KBIR approach to information retrieval shows a lot of potential. By using a controlled vocabulary to represent documents and queries, problems posed by natural language ambiguity can be overcome. The richness of natural language can be addressed by employing conceptual distance functions in place of pattern matching as the comparison process. There is no reason why the knowledge based approach could not be applied to the processing of non-textual information. As with the linguistic approach, once the information is described in natural language, perhaps using an interactive session with a knowledgeable user, the process of retrieval is the same as with textual information. Unfortunately, the effectiveness of knowledge based systems depends critically on the quality of the KB,

and yet systems to-date have not used a truly domain independent semantically rich KB.

2.2 What we Propose

The problem being tackled in our research, once again, is the domain independent semantic processing of information, where information is not necessarily defined as being textual, (but is captioned or described interactively). In the light of this objective, it is quite clear that a traditional approach would be ill advised and a linguistic approach would quite possibly produce bad results. The statistical approach proposes a non-semantic view of information and requires the information to be textual. Current linguistic approaches are also non-semantic and, although it would be possible to apply linguistic techniques to non-textual information, the description of information items would necessarily be far bigger than say an equivalent description in a knowledge based system. This leaves us with the knowledge based approach. However, as was pointed out in the previous Section, in order for the knowledge based approach to be successfully applied to any application one needs a good quality knowledge base. In our situation a good quality knowledge base is one which would include the following characteristics :

- *Domain independence*
- *Good coverage of different senses of polysemous concepts.*
This would enable us to address the problems posed by natural language ambiguity among word senses.
- *Semantically rich.* This would allow us to develop a sophisticated semantic similarity function to replace pattern matching between terms as the information comparison mechanism.

- *Include phrases.* This addresses the points raised with regard to multi-word phrases, (see Section 2.1 and 2.2).

We believe such a knowledge base can be constructed from the WordNet lexical database being developed at Princeton University. We thus propose to adopt the knowledge based approach. WordNet will be used as the knowledge base to represent both queries and information items from the body of information being processed. A semantic similarity estimator using the semantic knowledge encoded in WordNet will be developed to replace pattern matching as the information comparison process.

Bright et. al, in a very recent article, [Brig94], report on a system referred to as SSM (Summary Schemas Model), which was developed to address the issue of locating data in federated database systems, (refer to Section 1.2). As in our approach to the problem, they employed many of the techniques associated with KB information retrieval. Rogets thesaurus is used as a knowledge base and a simplified version of Rada's conceptual distance function as the comparison mechanism. Although an evaluation of the system, in terms of whether it actually "works" is not reported they say of Rogets thesaurus :

"Rogets provides only the most basic semantic relations, and the vocabulary is somewhat dated. In particular, Roget's hypernym (IS-A), links are not as meaningful as they could be ... ", [Brig94],

and they conclude of their conceptual distance function :

"SDM (Semantic Distance Measure), sophistication is currently limited by the available systems taxonomy. A more complex taxonomy (with more linguistic information), would allow more variation and control over SDM calculations. The SDM function is the core function that applies the power of the system taxonomy to provide semantically meaningful results to users... ", [Brig94].

We believe a knowledge base derived from WordNet, will not suffer from these problems encountered by SSM in its use of Roget's thesaurus as its knowledge base.

2.3 Summary

In this Chapter we reviewed the current state of related work in information retrieval. Having reviewed statistical, linguistic, and knowledge based approaches to information retrieval we opted to pursue the knowledge based approach. In particular we proposed the use of the lexical database WordNet as our knowledge base and the use of a semantic similarity function as our information comparison mechanism.

The remainder of this thesis is organised as follows. In the following Chapter we describe WordNet in much greater detail. In Chapter 4 we explain how WordNet was changed and expanded to become the knowledge base in our domain independent semantic information processing system. In Chapter 5 we derive a semantic similarity function. Possible applications of our system are discussed in Chapter 6 and we choose one of these applications to evaluate our approach to information processing. Chapters 7 and 8 present further details of the evaluation application and what work was involved in tailoring our system for the specific task. In Chapter 9 results of the application of our system to the evaluation application are presented. Our system is compared against another, non-semantic approach to the same application. Finally conclusions and suggestions for future research are presented in Chapter 10.

Chapter 3 - WordNet

3. Introduction

In this Chapter there is a description of the lexical database, WordNet. As will be seen in later Chapters, WordNet forms a large input to our research, and as such it is important to understand what it is and to introduce and explain the terminology used in its description at this early stage in the thesis. The discussion here focuses on the content and organisation of information within WordNet. For a discussion of the psycholinguistic theories behind the decisions made in its construction refer to [Mill90a, Mill90b, Felb90a, Felb90b and Beck92]. Also, a detailed explanation of exactly how WordNet is used in our research is deferred to later Chapters. For now it is sufficient to say that WordNet is the basic building block, supplying all the lexical knowledge, for our knowledge base. This knowledge base, KB, is extended in Chapter 4 and is used in conjunction with semantic similarity estimators, (developed in Chapter 5), in an information processing application, introduced in Chapter 6. The KB itself is used as a kind of controlled vocabulary to represent information in the information processing task. In an information retrieval task terms from the KB would be used to represent both the information request and the corpus of information to be searched. The semantic similarity estimators in such an application would be used to compare the representation of the information request against the individual KB representations of the information corpus to retrieve relevant items of information.

The remainder of this Chapter is organised as follows. In Section 3.1 there is an overview of the original motivating factors for the development of WordNet. The discussion goes on to describe the type and extent of information contained in WordNet. The physical and semantic organisation of this information is the central theme of this discussion. In Section 3.2 there is a description of the software made available with the WordNet package. In Section 3.3 there is a brief overview of the

current applications of WordNet. Future directions for WordNet are discussed in Section 3.4. Finally, conclusions of this Chapter are presented in Section 3.5.

3.1 WordNet

WordNet⁶ is the product of an ongoing research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English. The system has the power of both an on-line thesaurus and an on-line dictionary, and much more. In its simplest interpretation WordNet could be seen as a semantically organised dictionary. Traditional dictionaries organise information alphabetically. However, WordNet aspires to innovation in attempting to arrange its information by semantic meaning. This endeavour goes far beyond a traditional thesaurus where *synonym-of* is the only semantic relationship present. In WordNet use is made of the following lexical and semantic relations to arrive at a semantic organisation of concepts:

- | | |
|------------------------|----------------|
| - Synonym | - Entailment |
| - Hyponym (IS-A) | - Troponym |
| - Meronym (PART-WHOLE) | - Derived from |
| - Antonym | - Cause |
| - Attribute | - Similar |
| - Also see. | |

A relationship is lexical if it holds between word forms⁷, (e.g. synonym and antonym), and semantic if it holds between word meanings, (e.g. hyponym and meronym). A word form in WordNet is either a single word or a string of individual

⁶ WordNet is a public domain product. The developers wish to promote its use in as many applications as possible.

⁷ A word form is defined as the orthographic representation of a word, in other words the printed form without meaning.

words joined with underscore characters. These word strings are referred to as *collocations* and typically represent a single concept such as *fountain_pen*. WordNet currently holds approximately 95,600 different word forms, 51,500 of these are single words and the remainder are collocations. Due to the large number of polysemous words in the English language, many of these word forms are the same, (e.g. *bank* as in a river bank and *bank* as in a commercial bank). The different semantic meanings of these homonyms are represented by word forms made up of the same characters appearing in different places in WordNet's semantic network of concepts.

Word forms are themselves organised around logical groupings called synsets. Each synset consists of a list of synonymous word forms and relational pointers that describe relationships between the current synset and other synsets. The synonymy of word forms is not determined using the strict substitutability definition of synonymy whereby ‘..two expressions are synonymous if the substitutability of one for the other never changes the value of a sentence in which a substitution is made’, [Mill90a]. Instead synonymy is made relative to context, ‘.. two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value. For example, the substitution of *board* for *plank* will seldom alter the truth value in carpentry contexts, although there are other contexts of *board* where that substitution would be totally inappropriate.’, [Mill90a].

WordNet handles different syntactic categories by modelling words from each syntactic category in separate organising structures. All noun word forms are in one structure, verbs are in another, and adjectives and adverbs are in a third. In this research only the noun portion of WordNet is used and thus we concentrate on this portion of WordNet in our description. For a discussion on the organisation of verbs and adjectives in WordNet, refer to [Felb90a] and [Felb90b]. The remainder of this Section is organised as follows. In Section 3.1.1 we discuss the semantic organisation of information pertaining to noun word forms in WordNet. In Section 3.1.2 the physical organisation of this information is discussed.

3.1.1 Semantic organisation of Nouns in WordNet

WordNet contains approximately 57,000 noun word forms organised into approximately⁸ 48,800 synsets. Just under half of these word forms are collocations. The remainder effectively cover all nouns that appear in a high quality handheld dictionary. The relational pointers which semantically organise these concepts are :

- Hyponym/Hypernym (IS-A)
- Meronym/Holonym (PART-WHOLE)
- Antonym
- Attribute

However, of these the main organising relation is the hyponym or IS-A relation. This produces a hierarchical or tree like structure with generic concepts at the top and specific concepts near the roots. The resulting structure is hierarchical because most synsets are *a-kind-of* only one thing, for instance a { tree }⁹ is a kind of { plant } or a { car automobile } is a kind-of { motor_vehicle }, etc., however, like everything else there are exceptions. For instance a { rim } is both a kind of { boundary edge bound } and a kind of { round_shape }. However, these exceptions are relatively rare and for the discussion here the fact the structure is not strictly hierarchical is not of importance¹⁰.

From Figure 3-1 we can see how the concepts of a *car* and a *boat* are represented in WordNet. The edges in this diagram are not directed and this is to reflect the fact that many of WordNet's relational pointers are reflexive. The inverse of the hyponym relation is the hypernym or HAS-KIND relation. An important point illustrated by Figure 3-1 is the inheritance property of WordNet's hyponym link. The synset containing *vehicle* has features such as the part *splashboard* and this feature is inherited by its children synsets, { vessel } and { motor_vehicle }. These synsets in

⁸ Figures are approximate because WordNet is constantly growing.

⁹ Throughout this thesis a synset is represented by one or more word forms within the curly brackets, '{' and '}'.

¹⁰ We will, however, see that this situation does complicate the computation of semantic similarity in Section 5.2.1.

turn have their own distinguishing features which are inherited by their children. For instance, the synset { motor vehicle } has, among others, the parts *steering wheel*, *brake system*, and *fuel system* yet these features are not shared by its sibling synset { vessel }. Similarly, features such as *anchor*, *rudder*, and *hull* are unique to the synset { vessel } and its children synset. The child synsets of { motor_vehicle } and { vessel }, { car automobile } and { boat } respectively, inherit both the features of their parent synsets and the *splashboard* feature of their common grandparent synset.¹¹

Distinguishing features in WordNet are represented by the meronym or PART-WHOLE relation. [Win87] enumerate six different PART-WHOLE relations : component-object (branch/tree), member-collection (tree/forest), portion-mass (slice/cake), stuff-object (flesh/body), feature-activity (typing/programming), and place-area (Dublin/Ireland). In WordNet there are three meronym relations, PART-OF MEMBER-OF, and SUBSTANCE-OF. These equate to Winston's component-object, member-collection, and stuff-object relations. As with the hyponym relation, the inverse relations, (referred to as holonyms), are also present in WordNet; HAS-PART, HAS-MEMBER, and HAS-SUBSTANCE. The PART-OF meronym relation is by far the most widespread of these relational pointers. These relations can be thought of as traversing the hierarchical structure created by the hyponym relational pointers. The structure created by synsets connected by meronym relations tends to be non-hierarchical.

¹¹ Inheritance of this kind is one of the principles of the object oriented paradigm in computer science - see [Boo91].

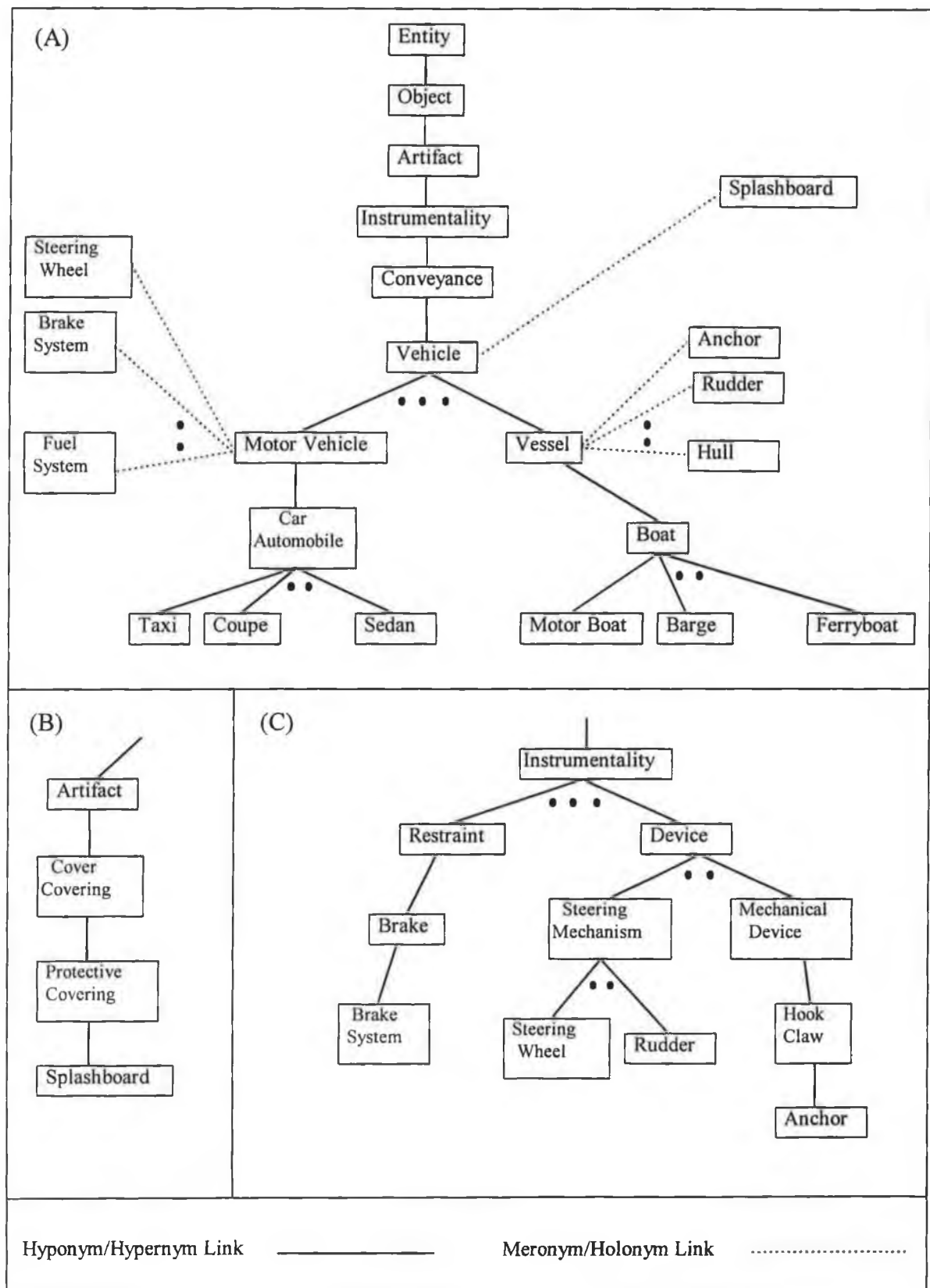


Figure 3-1 Extract from the WordNet Lexical Inheritance System

It is quite common for concepts connected by meronym relations to have multiple parents and children. For instance, the synset containing *point* is a meronym of *arrow*, *awl*, *dagger*, *fishhook*, *icepick*, *knife*, *needle*, *pencil*, *pin*, *sword*, and *tine*. Meronym relations also tend not to involve many parent-child levels, for example spoke is *part-of* a wheel which is *part-of* a car. If these were not regulated there would be a danger of the length of individual relations getting out of hand. If one takes the concept *atom* as an example, it is quite valid for all substances to be reduced down to this component element. To avoid this, ‘... the dissection of an object terminates at the point where the parts no longer serve to distinguish this object from others with which it might be confused...’, [Mill90a].

The remaining relational pointers, antonym and attribute do not play a central role in the organisation of nouns in WordNet. Antonym relations are most commonly found between nouns derived from antonymous adjectives. For instance *happiness* and *unhappiness* are antonyms and are derived from the adjectives *happy* and *unhappy*. The antonym relation is self reflexive so if A is an antonym of B then B is an antonym of A. The attribute semantic relation links noun attributes and the adjectives expressing their values. As with other WordNet relations the attribute relation is reflexive; the inverse relation could be said to relate adjectives and the nouns for which they express values. Examples would include:

Noun Synset	Adjective Cluster
Temperature	Hot / Cold
Pride	Proud / Humble
Age	Old / New
	Old / Young
Size	Big / Little
	Large / Small
Length	Long / Short

This is the first relation to cross from one syntactic category to another.

Neither the antonym nor attribute link types are widespread¹². As we shall see later, the antonym relation is not needed in our use of WordNet information, both because of the nature of the relation and the type of concepts connected by the relation. The attribute relation is particularly rare in WordNet and its inclusion in our research would involve the considerable overhead of including the adjective portion of WordNet. For these reasons, the attribute and antonym relations were not used in our research.

3.1.2 Physical organisation of information in WordNet

Information in WordNet is physically organised in flat files, one for each syntactic category, which are both human and machine readable. Synsets appear as individual records terminated by a line feed character and relational pointers point to the byte offset of the start of synset records in the data file. The fields within a synset record are as follows :

Byte_offset file_# pos syn_cnt { synset_details } ptr_cnt { ptr_details } | Glossary

The *Byte_offset* is an eight digit decimal integer indicating the data file byte offset of the start of this synset record, the *file_#* field refers to a lexicographer file used in the construction of the WordNet database, the *pos* field indicates the part of speech (always 'n' for noun in the noun data file), the *syn_cnt* field indicates the number of word forms in the synset. The *synset_details* field is repeated *syn_cnt* times and is made up of the two fields *word_form* and *sense_number*. The *word_form* field is simply the word_form, either a collocation or a single term, and the sense number field is used to uniquely identify homographs. The *ptr_cnt* field is a count of the number of relational pointers emerging from or connecting to this synset. The *ptr_details* field is made up of a pointer symbol followed by a space, followed by the byte offset of the target synset, followed by a space and a part-of-speech character

¹² A small program was written to discover just how many of the 48,800 WordNet noun synsets had an attribute link emerging from it. Only 80 synsets were found to contain 'attribute' links.

indicating which data file this pointer indexes into, followed by a space and a four digit hexadecimal from/to field. This from/to field is used by lexical relations to indicate the words in the source and target synsets that are linked by that relation. The pointer symbols used to encode relational pointers are illustrated in Table 3.1. The final field is an optional glossary that is sometimes¹³ included in the synset record. This field is made up of a short text explaining the meaning of the word forms within the synset. The developers of WordNet it necessary in many cases where synonyms alone were unable to differentiate between the fine sense distinctions made in WordNet.

Pointer Symbol	Relational Pointer
!	Antonym
@	Hypernym
~	Hyponym
#m	Member Meronym
#p	Part Meronym
#s	Substance Meronym
%m	Member Holonym
%p	Part Holonym
%s	Substance Holonym
=	Attribute

Table 3-1 Encoded Relational Pointers

In order to find a particular word form in the data file all word forms are listed in a large index file along with the byte offset of the synset within which they occur. The exact format of this index file is as follows :

word pos poly_cnt ptr_cnt [ptr_types] synset_cnt synset_offset [synset_offset]

The first field is the word form and the second is the part of speech indicator, again always 'n' for noun in the noun index file. The *poly_cnt* field is known as the index of familiarity and, as its name would suggest, is an indicator of how familiar the word form is in everyday discourse. Familiarity of a word form is estimated by its polysemy which is itself determined by counting the number of noun, verb, adjective or adverb senses the word has in the Collins Dictionary of the English Language. A

¹³ 67% of synsets have a glossary in WordNet version 1.4.

familiarity value of 0 indicates the word form doesn't appear in the dictionary and a value of 1 would indicate the term is probably quite technical in nature, see Table 3.2. The usefulness of the index of familiarity field would be greatly enhanced if it applied to word senses as opposed to word forms. In its present form the same index of familiarity is given to all senses of a word form, while in fact, some senses are used much more frequently than others.

Word Form	Polysemy
bronco	1
@-> mustang	1
@-> pony	5
@-> horse	14
@-> equine	0
@-> odd-toed ungulate	0
@-> placental mammal	0
@-> mammal	1
@-> vertebrate	1
@-> chordate	1
@-> animal	4
@-> organism	2
@-> entity	3

Table 3-2 Hypernyms of bronco and their familiarity values, taken from [Beck93].

For example, from the WordNet senses of *horse* given below, it is quite obvious that the fifth sense of horse is by far the most commonly used.

Sense 1

sawhorse, horse, sawbuck, buck -- (a framework for holding wood that is being sawed)

=> *framework, frame*

Sense 2

knight, horse

=> *chessman, chess piece -- (16 white and 16 black pieces)*

Sense 3

horse

=> *gymnastic apparatus, exerciser*

Sense 4

heroin, diacetyl morphine, H, horse, junk, scag, shit, smack -- (a morphine derivative)

=> *hard drug -- (a drug that is considered relatively strong)*

Sense 5

horse, Equus caballus -- (solid-hoofed herbivorous quadruped domesticated since prehistoric times)

=> *equine, equid -- (hoofed mammals having slender legs and a flat coat with a narrow mane along the back of the neck)*

Figure 3-2 Senses of horse in WordNet

The fourth field in the index file is a count of the number of different pointer types in all synsets containing the word form. The *ptr_types* field is a space-separated list of *ptr_cnt* different relational pointer codes. The *synset_cnt* field indicates the number of synsets the word form appears in and finally, the *synset_offset* field is a list of one or more byte offsets or indices into the corresponding data file for each synset in which the word form appears.

3.2 The WordNet Package

WordNet is currently on release 1.4, and the database is close to 13.5 megabytes in size. Standard with the database is a simple user interface application which provides full access to the WordNet database¹⁴. A second application, known as Escort, applies to the semantic concordance package which also comes as standard with the latest version of WordNet. Discussion of the semantic concordance package is deferred to Section 3.4. In this Section we concentrate on what software is available for users wishing to program their own interface to WordNet or gain access to the WordNet database from another application. This software can be categorised as either comprising of database search routines or morphological processing routines.

The search software accepts as input a word form and a search type variable. Searches are available for all relational pointer types and in both directions for reflexive pointers. Searches are by their nature recursive, however, given the large size of the search space it may be advisable to re-code them iteratively¹⁵.

¹⁴ There is both a graphical and command line version of the interface. The software is supported on six different platforms; Sun-4, RS6000, DECStation, NeXT, PC and Macintosh.

¹⁵ This is particularly true on machines with limited memory. In general, however, when the search returns with 'Search too large, try to narrow', the problem is that the buffer allocated to hold the result of the search is too small.

The software to perform morphological processing is necessary because only base forms of words are stored in WordNet. Two procedures are involved in the process of converting a concept into a form that is found in WordNet. Firstly there are lists of inflectional endings that can be detached from individual words to arrive at their base forms:

Noun	
Affix	Ending
s	
ses	s
xes	x
zes	z
ches	ch
shes	sh

Table 3-3 Inflectional endings for Nouns

Secondly, for words that are not regular and consequently can not be processed in an algorithmic manner, there exists an exception list. Each line of this exception list consists of an inflected form of a word followed by one or more base forms of the word. This list is kept in alphabetical order, thus permitting fast access using a binary search. Since some concepts, such as 'axes', have more than one base form, (axis and axe), the lookup to the exception list works as follows. On the first call it returns a specific base form. and on subsequent calls it returns any other base forms.

Collocations and single term word forms are processed differently. In the case of single word concepts the procedure is quite simple. The first step is to check for the word in the exception list. If it is found then the first base form is returned. Subsequent lookups return alternative base forms, if they exist. If the word is not found in the exception list then the algorithmic process that looks for a matching suffix is applied. If the matching suffix is found then the corresponding ending is applied.

In general only base forms of words, even those comprising collocation concepts such as "Object-Oriented_Database", are stored in WordNet. As such transforming the collocation "Object-Oriented_Databases" is then simply a matter of finding the base forms of the individual words which make up the collocation. Therefore non-conforming collocations such as "Customs_Duty" are entered in the exception list.

3.3 Current Uses of WordNet

WordNet has been around for a number of years now, however, the number of research projects using WordNet is quite small. In a recent survey carried out by the developers of WordNet to determine where its being used they had a total of just 47 responses. Of these only 33 had done more than download WordNet. These responses cited the following applications of WordNet¹⁶ :

- Word Sense Disambiguation
- Selectional preferences/Co-occurrences
- Machine Translation
- Language Learning
- Anaphora Resolution
- Query expansion in Information Retrieval
- Thesaurus
- Knowledge acquisition in NLP systems
- Study conceptual change/analogy
- Generate stimuli for studies of hemispheric lexical organisation.

¹⁶ Applications relevant to this research will be referred to as the topics of these applications are discussed in the thesis. References for the other applications can be found from the WordNet mailing list.

From the above it is clear that the applications of WordNet are numerous and varied. As mentioned previously, however, the total number of individuals and research teams in these areas, using WordNet is surprisingly low. One can only speculate as to the reasons for this reticence; for instance :

- Engineering Overhead Involved :

Any research team using WordNet would necessarily need a strong programming component. As was seen in Section 3.2 the software made available with WordNet is quite general and low level. WordNet was developed by psycholinguists but will find much of its application in the computational field of artificial intelligence. The resulting research projects will require researchers from many different disciplines to work very much closer than they would have been accustomed to in the past.

- Computational Power :

WordNet is a disk-based system and any application of WordNet in the large scale would require large powerful computers that can make many disk accesses very quickly

- Difficulty in making changes :

As we will see in the following Chapter, it is very difficult to add to, delete from or modify the WordNet database. This is due to its physical organising structure, a flat filed database. Any changes to information in these files immediately invalidates all byte offset references elsewhere in the database.

- Generality, doesn't suit any one domain :

WordNet was not built with any one application in mind and certain elements of it don't suit particular applications but may be invaluable for others. For instance, the existence of such fine sense distinctions in WordNet tend to be a hindrance to its application to most information retrieval tasks, however, such fine sense distinction makes it particularly suitable to its application in

language learning and as a thesaurus. Unfortunately, however, as the previous point makes, changing WordNet is not a straight forward task.

- Manual construction, questions of subjectiveness

WordNet is the first truly domain independent attempt to semantically organise information which has actually tried to deliver something and as such there are bound to be criticisms of various aspects of its approach. Many researchers query the subjectiveness of its manual construction, others criticise WordNet for not having a topical organisation of information, and still others question the choices made in the categorising or organising of nodes in WordNet.

- Waiting for someone else to take the plunge

As with most new ideas, people tend to prefer to wait until it is tried and tested before investing time and effort. Unfortunately a catch 22 situation arises in so far as a certain number of researchers must take the plunge and produce results of their research in order for WordNet to be deemed a suitable approach.

Despite these problems, WordNet remains a tremendous, unparalleled on-line store of information. It is slowly gaining acceptance within the sceptical research communities of many different disciplines and it is our opinion that it is only a matter of time before it will play a central role in many research endeavours.

3.4 WordNet the Future

As was mentioned in Section 3.1, WordNet is an ongoing project. In this Section we will give a brief outline of what features one may expect to find in future

releases of WordNet. Although WordNet is constantly increasing in coverage, primary future developments will be in terms of additional semantics. The two main strategies for this are firstly, to increase the number and types of relational pointers and secondly, to add context. In the following two subsections each one of these extensions is discussed in more detail.

3.4.1 Addition of Relational Pointers

The semantic richness of WordNet depends both on the number of relational pointers and on the number of types of relational pointers. At present there are approximately 126,500 relational links in WordNet and a link can be an instance of one of the 11 different types listed in Section 3.1. The number of instances of links is constantly growing and in this sense each new release of WordNet is semantically richer than the previous one. However, our interest in this Section is in what new types of links will be made available in the future.

In relation to nouns, Miller [Mill90b], identifies three distinguishing features for concepts :

- (a) Attributes
- (b) Parts
- (c) Functions

He uses the example of the concept of a *canary*. A canary would appear in an hypernym inheritance hierarchy under such concepts as *canary @-> bird @-> life_form*. Features one would expect of a canary would include parts such as a *beak* and *wings*, attributes such as *small* and *yellow*, and functions such as *can fly* and *can sing*. Notice that attributes are given by adjectives, parts are given by nouns, and functions are given by verbs. At present the only features encoded in WordNet are the parts features, (these are encoded using the three meronym/holonym pointers). Future

releases of WordNet will be expected to include relational pointers to encode the distinguishing features of attributes and verbs. Although the *attribute* link type already exists it currently only applies to adjectival nouns and, as was stated in Section 3.1.1, there are still very few instances of this link type. The fact that a canary is small can be represented by the existence of a non-reflexive attribute link between *canary* in the noun hierarchy and *small* in the adjective data file. The link is non-reflexive because although a canary is small, attributes are relative terms and when asked to list small things one is unlikely to include a canary. Furthermore, attribute links will be qualified by the immediate hypernym link so for our example, the attribute link can be interpreted as 'a canary is small for a bird'. There is currently no link between WordNet nouns and verbs. A new function/predicate pointer type which will connect WordNet verbs and nouns is currently under consideration, [Mill93].

3.4.2 Adding Context

New with the latest version of WordNet is a semantic concordance package. A semantic concordance is a '...textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon...', [Mill94]. In this instance WordNet is the lexicon and the Brown corpus is the text corpus. Thus far 103 files¹⁷ of the Brown corpus have been manually tagged with the appropriate noun, verb, adverb and adjective synsets in WordNet.¹⁸ At present, the only application using the semantic concordance is a utility called Escort. This X-Windows application accepts words and returns sentences from the corpus in which it occurs. It also allows the user to specify particular senses of the word and to find words that co-occur in a sentence. However, other uses and possible future developments envisaged as a result include :

¹⁷ Each of the files are approximately 2,000 words long.

¹⁸ Before being semantically tagged the text is syntactically tagged by a part of speech tagger developed by Eric Brill, [Bril93].

- Instruction: where someone unfamiliar with English will be able to get contexts for particular meanings of a given word. This is something that is generally not available in most dictionaries.
- Sense Frequencies: this would provide data on the frequency of occurrence of individual senses of words. Thus far this data has been largely unavailable. This could lead to a reformulation of the index of familiarity field in the WordNet database.
- Sense Co-occurrence: Information on sense co-occurrences will help to develop a topical organisation of concepts, something which WordNet is frequently criticised for lacking. Using sense co-occurrence information it will be possible to organise topically a concept such as *horse_racing* with concepts such as *horse*, *race_track*, *betting*, *bookie*, *trainer*, etc.. Obviously sense resolution will be facilitated by such an organisation.

The sense frequencies and co-occurrence information will obviously be very useful in automatic sense disambiguation and other applications of WordNet. However, these data build up slowly and are found to 'depend critically on the subject matter of the corpus being used', [Mill94].

3.5 Summary

In this Chapter we have described the lexical database WordNet. In Section 3.1 we described the semantic and physical organisation of information in WordNet. The discussion here concentrated on the noun portion of WordNet since this is the only part being used in this research. Subsequent Sections outlined the current uses of

WordNet and speculated on future directions. Many of the future developments are long term goals of the WordNet developers and it is unlikely that these extensions to WordNet will take place in the near future. It is more likely that individual elements of these extensions will form the basis of future releases of WordNet thus slowly achieving the larger, more ambitious goals.

In the following Chapter we describe how we use WordNet to build a knowledge base for our semantic information processing system. As mentioned in Section 3.3, WordNet was not developed for any one application and modifications and/or extensions required for an application are left up to the developers of the application. In Chapter 4 we describe in detail how we modified the structure of WordNet's information and extended the database with new fields specific to our purposes.

geometric space. Figure 5-1 below shows the two-dimensional conceptual space obtained for animals, [Rips73]. As can be seen the greater the similarity between two animals, the shorter the distance between them. The output of multidimensional scaling can often help researchers identify the structure of a conceptual space. An interpretation of the dimensions in Figure 5-1 would be predacity and size. Animals at the bottom of the space are predators compared with those on the top, and the animals to the right are smaller than those on the left.

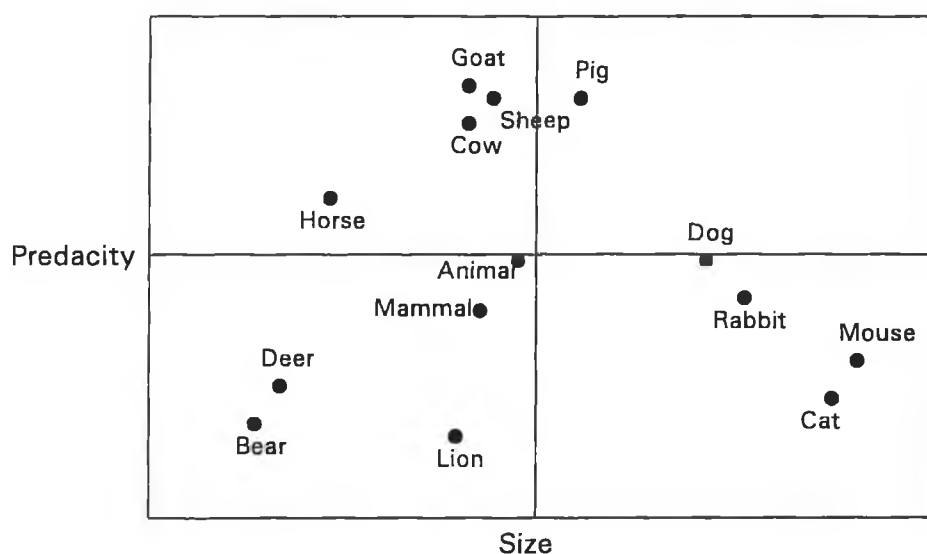


Figure 5-1 An example of similarity in a geometric space. (Taken from [Rips73])

The use of multidimensional scaling in general domains poses quite obvious problems, both in terms of interpreting dimensions and generating reliable proximity data. Besides the overall problems of geometric similarity models it is clear that the technique of multidimensional scaling is simply too complex to be automated by a machine. However, ‘..multidimensional scaling provides a useful technique for discovering structure in data across a wide range of basic and applied domains.’, [Bars92].

Simpler examples of geometric models of similarity can be seen in the conceptual distance similarity estimators reported in [Rada89, Kim90, Lee93, Gins93

and Shov85], and described in Section 2.2.3. These systems use hand built conceptual graphs and considerably simplify the process of estimating semantic similarity by constraining these graphs to have a single relationship type, (either an Is-A/Has_Kind-of or a Broader_Term/Narrower_Term link type). The concept graphs are quite small, relative to our KB,²⁸ and are generally domain specific. Semantic similarity between concepts is measured by aggregating the weights of links between concepts. This similarity estimator is applicable to our situation. Differences can be seen in the size, generality of domain, and number of relational link types of the concept graphs, but the very fact that our KB is a manually constructed semantic network advocates the use of a conceptual distance semantic estimator as a comparison mechanism. The extraordinarily large size of our concept graph effectively rules out the hand weighting of relational links. However, as was seen in Section 4.3, a mechanism to automatically weight relational links has already been developed. Also, the generality of the domain of our concept graph casts some doubt on the effectiveness of a conceptual distance similarity measure in our situation. However, we believe the existence of the non-hierarchical link types in our concept graph considerably adds to its semantic richness and by using these link types a reasonable approximation²⁹ of human judgement of conceptual similarity is possible. Refer to Section 5.2 for a complete discussion of our adapted conceptual distance similarity estimator.

The geometric model of similarity has been criticised by a number of researchers in the past, most notably by Tversky, [Tver77]. Problems with the geometric view of similarity are centred on the assumption that conceptual distance has metric properties. According to the metric property of minimality each concept in a concept space should be as similar to itself as any other concept is to itself and should be closer to itself than any other concept. Yet empirically gathered data violate these assumptions, [Tver77]. In an experiment to rate the similarity of letters it was found that humans were more likely to confuse some letters with another letter than be classified correctly (e.g. 'Q' is more likely to be called 'O' than 'Q'). Because these

²⁸ Allowing hand weighting of link types

²⁹ Obviously having 4 semantic link types as opposed to one will improve the KB's model of the real world, however, many other link types (e.g. attribute-of, function-of, etc., etc.), would be required to even approach a model of human memory.

letters are more similar to other letters than they are to themselves, they violate minimality. Similarly, according to the metric property of symmetry, the order in which people judge two concepts should not affect the distance between them. However, according to experiments carried out by Tversky, the choice of subject and referent in a similarity statement often determine the degree of similarity. For instance, the similarity ratings returned by humans for statements such as :

‘An ellipse is like a circle’

‘North Korea is like China’

were far greater than those returned for :

‘A circle is like an ellipse’

‘China is like North Korea’.

Finally, the triangular inequality property of a metric states the distance between any two concepts must be smaller than the sum of the distances from each to a third concept. However, Tversky again cites examples where this property is clearly being violated by human judgements of similarity. For example, according to triangular inequality, the conceptual distance between Jamaica and Russia must be smaller than the sum of the conceptual distances between Jamaica and Cuba and between Russia and Cuba. This is quite obviously not the case, since Cuba and Russia are quite similar politically and Cuba and Jamaica are similar geographically whereas Russia and Jamaica are not considered to be at all similar.

‘These problems for the geometric view of similarity suggest that people do not process similarity geometrically. If they did, their similarity judgements would not violate so many of its fundamental assumptions... ’, [Bars92].

However, these problems of the geometric model of similarity do not rule out our use of conceptual distance as a measure of semantic similarity. As stated earlier, we are not interested in a complete model of the human similarity measurement

process. Given the concept graph we are working with this would not be possible. In our situation, the metric properties of the conceptual distance estimator will not have a great effect on the large majority of estimates of similarity. Violations of minimality in the similarity ratings of printed words are not as likely as in the case of the broader view of similarity seen in the ratings of sounds, pictures and symbols. Violations of the symmetric property of a metric could be seen to be the exception more than the rule. Generally speaking there is not a great difference in the salience of concepts in a similarity judgement. The same could be said of the triangular inequality property, the examples cited by Tversky assume a complete model of the real world. In our semantically scaled down model of the world, with just Is-A, Member-of, Part-of and Substance-of relational links, examples of violations of the triangular inequality are more difficult to come by.

5.1.3 Contrast Model of Similarity

Based on the set-theoretic model of human memory, Tversky's contrast model simply represents concepts as feature lists, [Tvers77, Tvers78]. According to this model the similarity of two concepts is measured by counting the number of common and distinctive features of both concepts. Common features are properties that both concepts share such as *wheels* for *car* and *bicycle* and distinctive features are properties of one concept not shared by the other such as an *engine* for the pair *car* and *bicycle*. Tversky's model can be formally represented as follows :

$$Sim(x, y) = k_c \cdot F(C) - k_x \cdot F(D_x) - k_y \cdot F(D_y), \quad k_c, k_x, k_y \geq 0$$

where C represents the number of features common to X and Y, D_x represents the number of distinctive properties in X, D_y represents the number of distinctive properties in Y, and the function F weights particular features according to their salience in the particular similarity comparison under consideration. Finally, the constants k_y , k_x , and k_y weight the importance of common and distinctive properties in

particular comparison, taking on different values for judgements of similarity and dissimilarity.

The power of this model of similarity overcomes the problems caused by violations of minimality, symmetry, and triangular inequality. For minimality, concepts can vary in how similar they are to themselves, depending on their number of features. For asymmetry, the order in which people judge two concepts can affect their similarity if the values of k_x and k_y differ. If k_x is greater than k_y then similarity is always greater when the concept with the most distinctive properties comes second in the comparison. Taking the *China/North Korea* example we can see that China could have more distinctive properties than North Korea because people generally know more about China. If $k_x > k_y$ then the negative impact of both concepts' distinctive features on similarity is least when North Korea is first. Finally, violations of triangular inequality are handled by the contrast models common properties. Two concepts can be similar to a third concept for two different reasons and yet have neither property in common with each other. For example, Jamaica and Cuba have a common geographic property and Cuba and Russia have a common political property, yet neither of these properties are in common for Jamaica and Russia.

Tversky's contrast model has had wide acceptance. One of its major drawbacks, however, is to be found in the fact that an infinite number of properties are true of any concept. Thus a *car* can be yellow, it can be bought, and it can be left on the side of the road, however, these properties could also be true of a banana. Obviously as the number of common properties approach infinity, measures of similarity become meaningless and some method of arriving at salient properties for a comparison is necessary. Such a mechanism requires a complete model of the real world which is semantically rich enough to include contextual information in the comparison process. This allows the salient features of the concepts under consideration to come to the fore.

Applying Tversky's model of similarity to our situation would appear to be very difficult. Our WordNet derived KB is clearly semantically deficient in its model

of the real world. However, accepting our limitations it should be possible to take some of the ideas from Tversky's feature based model to come up with a second similarity estimator. Attributes of a concept in our KB could be said to be the set of subordinate terms below the synset it appears under in the hierarchical concept graph. The similarity of two concepts could then be estimated by the degree of overlap of each others subordinate terms. The information content value synsets could be used as a measure of this degree of overlap. Resnik, [Resn93a], describes a similarity estimator based on the information content value of the first synset in a HCG that subsumes the synsets of both concepts under consideration. Synsets near the top of HCGs tend to have broad meanings and as such small information content values. Therefore if, when looking for a subsuming synset for two concepts, we have to travel to near the top of the HCG, the two concepts are likely to have little in common and therefore be dissimilar. This is reflected in the small information content value of the subsuming concept. Correspondingly, if two concepts are quite similar the subsuming concept will be quite deep in the HCG and will have a larger information content value. Consult Section 5.2 for a complete discussion of this semantic similarity measure, (referred to as the information based or information theoretic similarity estimator).

An interesting parallel can be drawn between the information content of the subsuming concept in this similarity estimator and Tversky's definition of the salience of a feature. Tversky explains the salience of a feature by diagnosticity and intensity. Intensity refers to the factors that increase intensity; e.g. loudness of a noise, size of a shape, or frequency of an item, etc. Diagnosticity is context dependent and relates to the classificatory power of features. As was seen in Section 4.2, the two inputs to a synset's information content value are the frequency of occurrence of its member words and its relative position in the HCG, (since the frequency of occurrence of the member words of its subordinate synsets also contribute to information content value). Frequency of occurrence could be said to be a synset's intensity and its position in a HCG could be said to be its diagnosticity. This provides us with a further interpretation of the information based similarity estimator as an implementation of Tversky's contrast similarity model.

5.2 Similarity Estimators Employed

In the previous Section there was an overview of related research on similarity. From this overview we chose two measures of similarity that can be adapted for use in our semantic information processing system; the geometric based conceptual distance measure and the set theoretic related information theoretic measure. In the following subsections there is a more complete discussion of these estimators of conceptual similarity. For each measure there will be a description of how it operates, a discussion on how it was implemented and a brief discussion on its obvious strengths and weaknesses.

5.2.1 Conceptual Distance Similarity Estimator

The conceptual distance approach is based on the work of [Rada89, Kim90, and Lee93] and uses edge weights between adjacent nodes as an estimator of semantic similarity. According to Rada :

‘.. An edge weight between two adjacent nodes can be used as the measure of conceptual distance between two nodes, since it reflects the degree of relationship between the two nodes. For example, if edge weight W_{ij} is larger than W_{ik} , then the index term t_i is conceptually closer to the index term t_k than to index term t_j . Therefore, the sum of edge weights along the shortest path connecting two nodes reflects the conceptual distance between the two. ..’

It is assumed that the less the conceptual distance between two nodes becomes the more similar they become.

In our application of the conceptual distance measure of similarity we used the concept nodes and weighted links from the eleven HCGs of our WordNet derived KB. These concept graphs are considerably larger and not as domain specific as those used by Rada, Kim, Lee, Ginsberg or Shoval. As was seen in Section 4.3, a method of

automatically weighting relational links had to be developed to overcome the impossible problem of hand-weighting links. Difficulties posed by the generality of the domain of our HCGs should be alleviated by the added semantics introduced by the non-hierarchical link types. Conceptual distance is thus defined as the shortest path between the two concepts, taking the weights of links on the path into account and using links of any type.

Although this definition is quite simple its implementation proved a little more complex. Initially it was intended to compute the conceptual distance using spreading activation, as has been the approach of some other researchers in the area, however, due to the existence of weighted links, this was not possible. It was found that spreading activation can miss short paths when simply counting the links. For example, if $\text{Dist}(X,Y) = 1.0$, $\text{Dist}(X,Z) = 0.3$, and $\text{Dist}(Y,Z) = 0.5$, the shortest path between X and Y will traverse Z and have a total weight of 0.8, but spreading activation would settle for a path of length 1 but a weight of 1.0.

Following this an investigation was made into the idea of using a shortest path algorithm, such as Warshalls algorithm [Sedg88], to pre-compute the shortest paths between all HCG nodes and store them in a large look-up table. However, considering the fact HCGs contain of the order of tens of thousands nodes, the resulting look-up table would be excessively large. This approach also necessitated the recalculation of shortest paths whenever the HCG was expanded or the weighting scheme was altered in any way.

Shortest paths are thus approximated by looking at the common ancestors and descendants of concept nodes. The first step involves finding the ancestors of both concepts. This is accomplished by recursively tracing the links back to the root concept. If on this trace a node is found with more than one parent then each path is pursued. The path(s) to the root for each concept are then compared. If the paths intersect then a value for the conceptual distance can be computed. A check has also to be made for common descendants due to the existence of multiple inheritance, as Figure 5-2 illustrates. Assuming all links have a weight of one, the distance between

'K' and 'L' is calculated as being equal to four if only the common ancestors are taken into account, (darkened path). However, as can be seen from the graph, the actual distance is two because the common descendant 'P' allows the path 'K' -> 'P' -> 'L'. Comparing the descendants of two concepts nodes amounts to comparing the two subtrees rooted at these nodes in the HCG. The first step is to locate nodes with multiple parents in either of these subtrees. The set of multi-parented nodes from each subtree can then be compared. If there are any nodes that appear in both sets then it can be deduced that a path among the descendants exists.

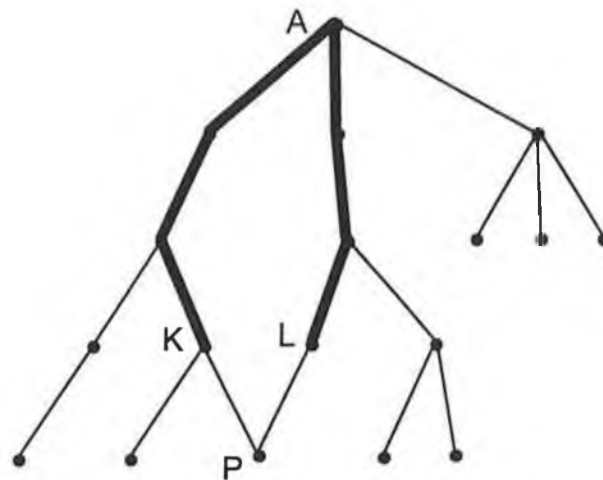


Figure 5-2 Multiple Inheritance

A final complication of the conceptual distance measure is with regard to the use of the non-hierarchical relational links. Non-hierarchical links in WordNet are organised so they are inherited from parent nodes in the IS-A hierarchy, as Figure 5-3 illustrates. The fact that that a railway car is part of a train is captured by the MEMBER-OF/HAS-MEMBER link between both concept nodes, however, the link is not replicated for the fact a railway car is also part of a freight train. Instead it is implied, through an inheritance property of the IS-A hierarchy, that given that a freight train is a train and a railway car is a member of a train then a railway car is also a member of a freight train. In order to take this fact into account it was necessary to follow the non-hierarchical links of parent nodes in the IS-A hierarchy when determining conceptual distance using the non-hierarchical relational links.

As can be seen from section 2.2.3, Rada used his conceptual distance measurement in a knowledge based information retrieval system to measure the conceptual distances between index terms of queries and documents. Through subsequent experiments they have shown that their algorithm 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.

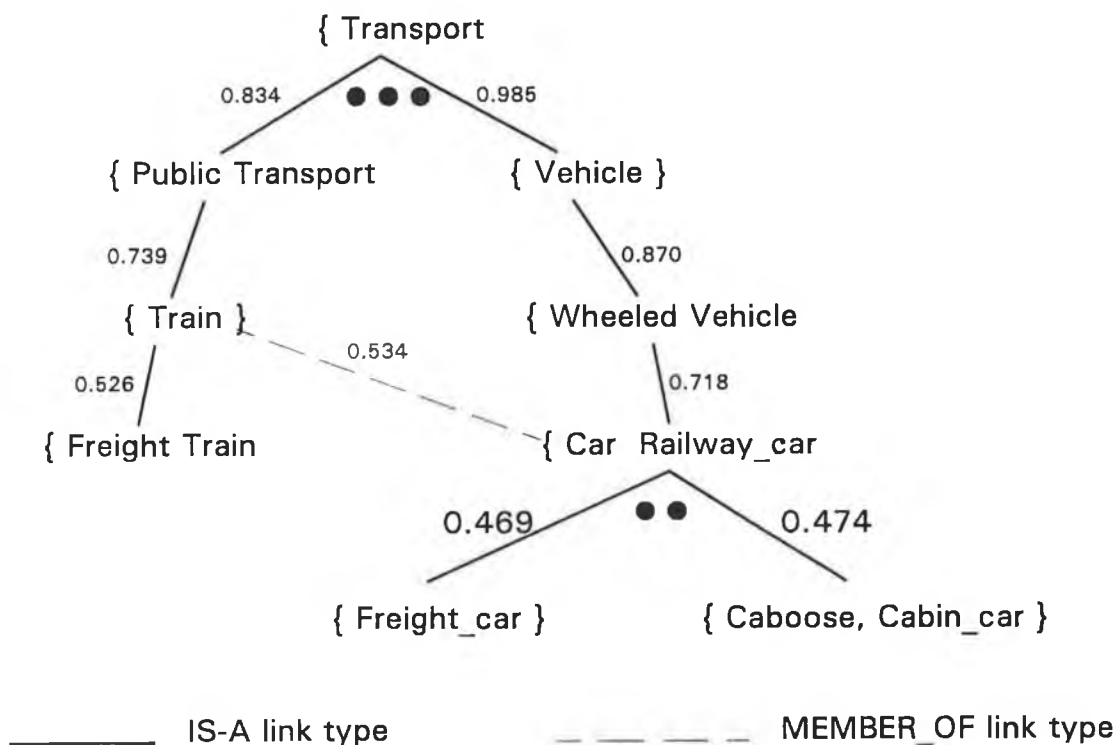


Figure 5-3 Weighted KB Extract

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 the HCG, 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. Refer to Section 8.2 for a more complete discussion on problems encountered with the conceptual distance similarity estimator.

5.2.2 Information Based Similarity Estimator

The information based approach to measure semantic similarity is based on work carried out by Philip Resnik, [Resn93a, Resn93b] and, as we demonstrated in Section 5.1.3, can be related to Tversky's contrast similarity model. As was explained in Section 4.2, Resnik views noun synsets as a class of words and 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 thus considered in terms of class similarity. The similarity therefore between two concepts is approximated by the information content of the first class in the noun hierarchy that subsumes the classes of both concepts, (see Section 4.2 for a discussion on the calculation of information content values). The similarity of two concepts can thus be expressed as :

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

where C_1 and C_2 are the classes of the input concepts, (C_i) is the set of classes dominating both C_1 and C_2 , and $\log \frac{1}{P(C_i)}$ is the information content of class C_i .

The method could probably be best illustrated by an example³⁰. If we assume we wish to discover the similarities between the following concepts : 'car', 'bicycle', 'banana', and 'fork'.

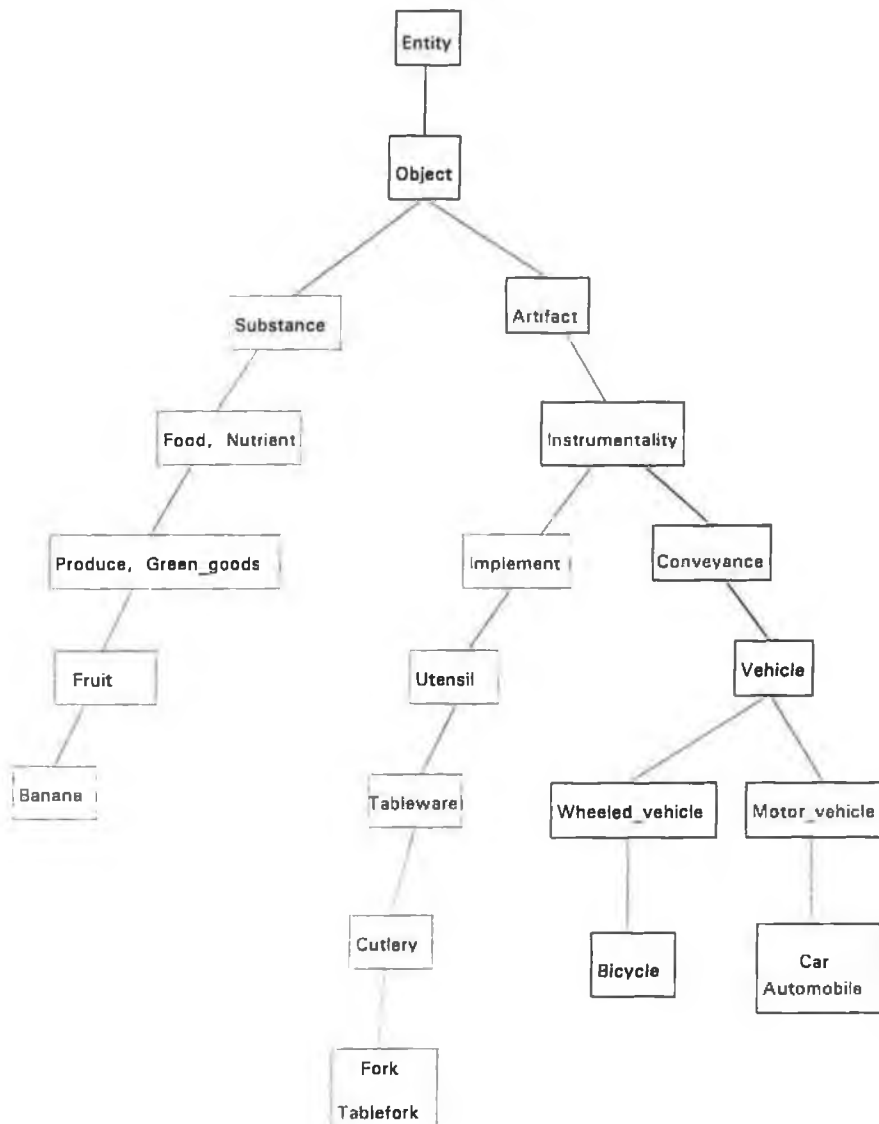


Figure 5-4 KB Extract for the concepts 'car', 'fork', 'bicycle', and 'banana'

Taking first $\text{Sim}(\text{car}, \text{bicycle})$, we see from Figure 5-4 that our KB has six classes to which both 'car' and 'bicycle' are subordinate :

³⁰ We use version 1.4 of WordNet to replicate one of Resnik's examples

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		

Cars and forks thus seem considerably less similar than cars and bicycles, however they are more similar than cars and bananas. This can be explained in the fact that forks and cars are objects that people use (instrumentality node), whereas all that can be said in terms of the similarity of cars and bananas is they are both nonliving things (object node).

The implementation details for this similarity estimator are quite straightforward. Having located the KB synsets of the concepts under consideration, a simple recursive trace back up the Is-A hierarchy to give a list of ancestral nodes is carried out for both concepts. These traces are then compared to find common ancestors. The information content value of this common ancestor is then used as a

measure of the semantic similarity between both concepts. The only complication can be seen in the existence of multi-parented nodes. Although the Is-A graph is, in general, strictly hierarchical, isolated examples of multiple inheritance means that certain nodes have more than one parent. In these situations there is more than one path to the root of the HCG for that node. As a result, all the root paths of a concept must be compared against those of the second concept when determining the information based estimate of semantic similarity.

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 hierarchy 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 meaning. This gives the same (exaggerated) information content value to a 'commercial bank' and 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. If one takes for example 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 particular 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, (refer to Chapter 7 for further discussion). 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 :

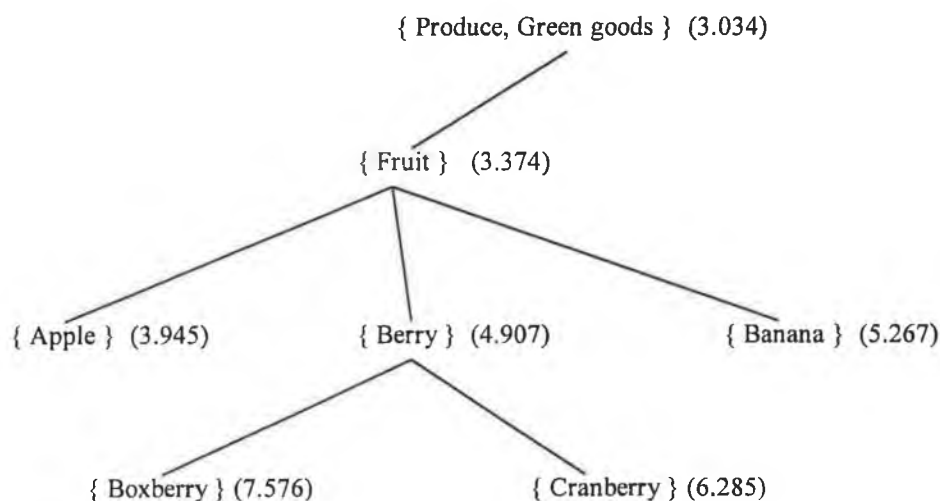


Figure 5-5 KB Extract showing violation of minimality in information based similarity estimator

Above is an extract from the KB, the numbers in brackets after the synsets are their information content values. From here 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. Also, the similarity between a boxberry and a cranberry is closer, (4.907), than the similarity between fruit

and itself, (3.374). This is a clear violation of the minimality property of a metric. The fact that the information based measure of similarity is non-metric relates back to its connection with Tversky's set-theoretic contrast model. However, as explained before, violations of minimality and the other metric properties are undesirable in our simplified model of the world and could be seen to have a bad effect on system performance.

5.2.3 Conclusions on Similarity Measures

In the previous Sections we described the semantic similarity estimators we intend to use in our information processing system. It is proposed to use these measures in place of direct pattern matching between words. In so doing it is believed many of the problems associated with using pattern matching as a comparison mechanism in the processing of information will be addressed.

The information theoretic and concept distance measures of similarity are quite different in their approaches to estimating semantic similarity and there is no obvious way of combining them to give a single unified measure of similarity. Between them they use all the information made available by the KB and in so doing provide the best similarity given the resources available. Although both measures are in some way derived from theories of the human similarity process, the emphasis here is on their use in a computational information processing task and not on their suitability as models for some human cognitive process.

In the following Section there is an initial evaluation of both similarity measures using human judgement as a baseline. As a result of these experiments it is hoped to show that the use of the information based and conceptual distance semantic measures with our KB are sufficiently accurate estimators of human judgements of conceptual similarity. A more complete evaluation of the KB and similarity estimators in an application of our system is presented in the following Chapters.

5.3 Psychological Evaluation

Resnik points out in his thesis "...there is not yet a standard way to evaluate computational measures of semantic similarity.", [Resn93a]. However, following a scan of the literature on this topic, it seems to us that the accepted baseline is human judgement,

'Semantic similarity is easily estimated by asking people to rate pairs of words with respect to their likeness of meaning' [Mill91].

Various studies, [Hen69], [Rub65], and [Mill91], have produced results that support the assumption that '..intelligent persons who know a language can reliably assess the semantic similarity of any two words they know how to use.', [Mill91]. As such, it seems the most obvious method of evaluating our estimators of semantic similarity is to use the judgement of humans as a baseline.

We used the results of a set of word-pair similarity tests described in [Rube65] and replicated in [Mill91] and [Resn93a]. Section 5.3.1 presents the test data set and describes the background to the experiments³¹ carried out by Rubenstein and Miller. Section 5.3.2 contains the results of using the information based and a basic configuration of the conceptual distance measure as semantic similarity estimators. In Section 5.3.3 various weighting strategies are evaluated in an attempt to arrive at the best link weighting mechanism for the conceptual distance estimator of semantic similarity. Through these experiments it is shown that our method of computing information content values improves upon Resnik's original implementation and that the link weighting mechanism proposed in Section 4.3 is an improvement on that of Sussna.

³¹ The conceptual distance basic configuration is using Sussna's link weighting mechanism.

5.3.1 Test data set

The data set originally used by Miller and Charles and subsequently by Resnik consists of the 30 noun pairs displayed in Table 5.1. The values show how subjects from the Miller and Charles' synonymy experiments rated the similarity of the noun pairs. The 5 point scale, 0 to 4, placed perfect synonymy at 4 and no similarity at 0.

Word Pairs		Miller and Charles
car	automobile	3.92
gem	jewel	3.84
journey	voyage	3.84
boy	lad	3.76
coast	shore	3.70
asylum	madhouse	3.61
magician	wizard	3.50
midday	noon	3.42
furnace	stove	3.11
food	fruit	3.08
bird	cock	3.05
bird	crane	2.97
tool	implement	2.95
brother	monk	2.82
crane	implement	1.68
lad	brother	1.66
journey	car	1.16
monk	oracle	1.10
cemetery	woodland	0.95
food	rooster	0.89
coast	hill	0.87
forest	graveyard	0.84
shore	woodland	0.63
monk	slave	0.55
coast	forest	0.42
lad	wizard	0.42
chord	smile	0.13
glass	magician	0.11
noon	string	0.08
rooster	voyage	0.08

Table 5-1 Average Human Similarity Scores for 30 Noun Pairs

The experiments correlated to a very high degree with a similar set of experiments, (using the same noun pairs), carried out by Rubenstein in 1965, [Rube65].

It is proposed to evaluate the information based and the conceptual distance estimators of similarity by using them as another subject in the evaluation of the semantic similarity of these noun pairs. In order for this to take place each of the nouns in the data set must be located in the knowledge base and in situations where there are a number of senses for a noun in the KB, an appropriate sense has to be agreed on. All of the nouns were found in the KB except the noun 'woodland'³², as such only 28 noun pairs are usable for the experiment. Appendix C illustrates the noun pairs that are ambiguous in WordNet and identifies the sense used in our experiments. These senses were not chosen in order to obtain the best results but rather to reasonably reflect what a human subject would choose to be the most likely sense, given the noun pair.

5.3.2 Initial Results

Initial experiments were carried out to evaluate our implementation of the information theoretic approach and to evaluate the basic configuration of the conceptual distance approach, (i.e. using Sussna's link weighting mechanism). Resnik had used the same data set to evaluate his information theoretic approach so there is an opportunity to compare both implementations. The conceptual distance approach has not been tested with this data set before.

Table 1 in appendix D shows the results of the initial experiment. The first column shows Miller and Charles' human similarity evaluations, the second column shows Resnik's results, the third column shows the results of our information based

³² Although code was developed that allows extensions to be made to the knowledge base, we chose not to add a synset containing 'Woodland'.

implementation, and the final column shows our conceptual distance results. The product moment correlation, [Lehm75], between Resnik's results and those of Miller and Charles is $r = 0.7677$. The same correlation for our information based implementation and Miller and Charles' was computed to be **0.8147**. This shows an obvious improvement in our implementation. This can be attributed both to the larger text corpus used to compute information content values and to the improvements introduced due to the handling of collocations.

With a resulting correlation coefficient of **-0.730** the conceptual distance similarity estimator was not as good the information based system. Before carrying out the experiments we had hypothesised that the conceptual distance estimator would surpass the information based one in situations where the concepts in question were connected by non-hierarchical links. Unfortunately, and very surprisingly, only one of the noun pairs in the test data set are connected by non-hierarchical links. The pair '*Furnace - Stove*' were found to have the common part '*grate*'. It is believe this absence of non-hierarchical links is coincidental for this small data set and not reflective of the KB as a whole.

However, even leaving aside the absence of non-hierarchical connectors for this test set, the conceptual distance estimator performed considerably poorer than the information based similarity estimator. This would point towards the need to improve the automatic link weighting mechanism. The following Section reviews our proposals for improving the weighting strategy from Section 4.3. Section 5.3.4 presents the results of these improvements.

5.3.3 Evaluation of Weighting Strategies

This Section presents a brief review of the automatic weighting mechanism used to weight relational links in our HCGs (for a more detailed discussion on the subject refer back to Section 4.3). It can be recalled from Section 4.3 that Sussna, in

[Suss93], hypothesised that the value for the weight of a link is affected by the following :

- (a) the density of the HCG at that point - distance is less in a dense part of the network
- (b) the depth in the HCG - distance shrinks as one descends a hierarchy

As well as altering the way in which local density and depth scaling are measured we proposed that a third factor should be taken into account in the automatic link weighting mechanism :

- (c) the strength of connotation between parent and child nodes.

Sussna's link weighting mechanism can be described as follows :

$$W(X, Y) = \frac{W(X \xrightarrow{r} Y) + W(Y \xrightarrow{r'} X)}{2d} \quad (1)$$

where X and Y are two adjacent nodes, $\xrightarrow{r}$ is a relation of type r, d is the depth of the deeper of the two nodes X and Y, and

$$W(X \xrightarrow{r} Y) = \max_r - \frac{(\max_r - \min_r)}{nr(X)} \quad (2)$$

Here, $\max_r$ and $\min_r$ are the maximum and minimum weights possible for a relation of type r, and nr is the number of relations of type r leaving node X or Y, (depending on whether we are looking at $w(X \xrightarrow{r} Y)$ or $w(Y \xrightarrow{r'} X)$). :

Under our proposal the local density of a link connecting a source node to a destination node is estimated by :

$$Den(X) = \frac{Sour_fan + (par_fan + sib_fan + des_fan) \frac{d}{d+1}}{Num_fans} \quad (3)$$

where d is the depth of the deeper of the source and destination nodes, $sour_fan$, par_fan , sib_fan and des_fan are the fanouts of link type r for the source, parent(s) of the source, sibling(s) of the source and the destination node respectively, and Num_fans is the number of the above fanouts present for a given source-destination pair. Each fanout is weighted according to its perceived importance. At present the fanout between the source and destination nodes is given a weight of 1.0 and all other fanouts are given a weight equivalent to $d/d+1$. This new density estimator replaces $nr(X)$ in (1).

In an attempt to overcome problems caused by the assumption that concepts at the same depth in a HCG are at the same level of abstraction, the depth scaling factor was changed to include the nodes information content value. Following experimentation with various combinations of information content values and the original depth scaling factor, the following formalism was decided upon :

$$Ds = ABS\left(\frac{info_cnt}{\delta} + 1\right) \quad (4)$$

where $info_cnt$ is the information content value of the deeper of the source and destination nodes of the link being weighted, and δ is the standard error from the mean for the information content values of the HCG in question.

Finally the formalism for calculating the strength of connotation between a source and destination node is represented as :

$$Sti = 1 - \frac{\sum \frac{1}{info_cnt}}{1} \quad (5)$$

where St_i is the strength of connotation of link i , $info_cnt_i$ is the information content of the destination synset and n is the total number of links of this type emanating from the source synset.

Having come up with improvements to the original weighting strategy it is now necessary to test each improvement individually and to decide on a unified automatic weighting formalism. The same data set as was used to evaluate Sussna's weighting strategy is used for this testing phase. The procedure was to isolate each component, as just identified, and experiment with them to determine their effect.

Dealing first with the strength of connotation factor, one way to include this in the original weighting strategy would be to multiply it by the local density estimator :

$$W_i(X, Y) = \frac{W(X \xrightarrow{r} Y)(St_i) + W(Y \xrightarrow{r'} X)(St_i')}{2d} \quad (6)$$

where $w(X \xrightarrow{r} Y)$, $w(X \xrightarrow{r'} Y)$, and d are as per (2), St_i and St_i' are the strengths of connotation of the link and its inverse respectively. It should be noted that expression (2) weights all the links of the same type between X and Y with the same undirected weight, however, expression (6) individually weights undirected edges between X and Y.

Table 2 in Appendix D shows the results of using equation (6) to automatically weight links. A very promising correlation coefficient of **-0.7704**³³ was calculated when these results were correlated against those of Miller and Charles' human subject averages. This shows a considerable improvement over the results of just using Sussna's weightings.

³³ It should be borne in mind that for the conceptual distance similarity estimator the smaller the similarity value the more similar the concepts are

The next step was to determine the effect of replacing Sussna's local density estimator with the one described by (3). The formalism in (6) remains the same except $w(X \xrightarrow{r} Y)$ and $w(X \xrightarrow{r'} Y)$ are altered to be :

$$W(X \xrightarrow{r} Y) = \max_r - \frac{(\max r - \min r)}{Den(X)} \quad (6)$$

where $Den(X)$ is as per (3). Again the results of using this formalism can be found in Table 2 of Appendix B. When these results were correlated against those of Miller and Charles a product moment correlation coefficient of **-0.7956** was recorded. This again indicates a slight improvement in results.

The final factor to be experimented with was the depth scaling. It was not at all obvious how we could introduce information content values into the depth scaling. Direct use of the information content values, as with the depth in a HCG, would not be expected to give good results³⁴. However, the results of using (4) as the depth scaling factor can be found in Table 2 of Appendix B. The correlation coefficient with the average of the human ratings was **-0.7820**. This is not as good a result as was obtained for scaling by dividing by the depth in the HCG, however, the result is not significantly worse. A final experiment involved combining both scaling methods :

$$W_i(X, Y) = \frac{W(X \xrightarrow{r} Y)(Sti) + W(Y \xrightarrow{r'} X)(Sti')}{\frac{d}{2} + Ds} \quad (7)$$

The result of this configuration was a correlation coefficient of **-0.7939**. This is the same as the result obtained for the original depth scaling, however, with a larger data set it is believed the inclusion of the information content values would produce even better results.

³⁴ An experiment testing the direct use of information content values proved this to be correct, the results were poorer than with the original depth scaling.

5.3.4 Conclusions on Psychological Evaluation

Important results from the tests presented in this Section are :

- The use of a WordNet derived KB and the information based and conceptual distance similarity measures are reasonable approximations for human similarity judgements
- The information content values calculated using the WSJ text corpus with special handling of collocations is an improvement on Resnik's original implementation
- The extensions to Sussna's automatic weighting scheme described in Section 4.2 lead to an improvement in the conceptual distance approach to estimating similarity.

The results here provide us with evidence that our approach to performing semantically based matching is promising as well as allowing us the opportunity to anchor certain variables in both semantic similarity estimators.

5.4 Conclusions

This Chapter highlighted the problems posed by the use of pattern matching in a retrieval system. The need for semantics in the comparison process was responded to by the introduction of two semantic similarity estimators. Section 5.2 presented a brief overview of related work on semantic similarity and our two similarity measures, the information theoretic and conceptual distance, were presented in terms of this previous work. In Section 5.3 both similarity measures were described in greater detail with discussions on how they operated, how they were implemented, and brief discussions on their strengths and weaknesses. Finally in Section 5.3 there was an

initial evaluation of our KB and similarity estimators, using human similarity judgements as a baseline.

Having reached the stage where we have built our KB and developed semantic similarity estimators Chapter 6 will present a sample of possible applications. Each application involves an information processing task in which we use the KB as a controlled vocabulary to represent a large volume of information. The semantic similarity estimators are used as the comparison mechanism in the processing of information. One of these applications will be chosen to evaluate our semantic knowledge based system.

Chapter 6 - Applications

6. Introduction

In the course of the previous two Chapters we have described some of the components to be used in a semantic information processing system. In Chapter 4 the KB was built using WordNet as the main deriving factor and in Chapter 5 the information based and conceptual distance semantic similarity estimators were proposed in place of pattern matching as the comparison process. In this Chapter a number of possible applications will be described and one of these applications will be chosen as our evaluating application. In all of these applications the KB is used as a controlled vocabulary and the similarity estimators are used as the comparison mechanisms.

Our system is not only applicable to information retrieval tasks, although this may seem the most obvious application. A variation on information retrieval known as information filtering is another possible application of our system. Unlike information retrieval where the query is matched against a fixed anthology of information, in information filtering the query is fixed and the information can be thought of as a dynamic stream which is matched against the query. Even more removed from a traditional information retrieval task one can perceive possible applications of our system in the processes of automatic text abstracting and the automatic construction of tours in hypertext systems, [Dunn93, Guin92, Niel89, and Smea90]. In the latter application users enter a query, (or perhaps more correctly a topic of interest), it is then the job of the system to construct a tour of the hypertext consisting of information nodes relevant to the user entered topic(s). Traditionally the method of determining the relevance of a node has been through direct pattern matching of terms of the user request with terms in the hypertext nodes. Obviously our system could be applied in this situation with the similarity estimators being used along with KB representations of the user request and hypertext nodes. A final example of an application not involving the retrieval of information could be

described as caption comparison, whereby captions of text are compared against each other to determine degrees of similarity. An interesting example of this type of application is currently being developed by the Garda Siochana. In their application they are investigating a way of automating the procedure of comparing witness statements to determine the degree of overlap. Again the witness statements could be represented by sets of terms in the KB and the semantic similarity estimators could be used as the comparison mechanism.

Although we have just discussed possible applications of our system in terms of whether they are retrieval or non-retrieval applications, for the purposes of the rest of this Chapter we categorise applications as being either Self-Describing (SD) or NonSelf-Describing, (NSD). All of the applications discussed above, document retrieval, information filtering, generating hypertext tours and caption comparison could all be categorised as SD applications. An SD application is characterised by the fact that it is possible to automate the procedure of generating a KB representation of the information collection. For all the applications above, the information collection is in the form of written text which can be processed to find KB terms representative of the text. Examples of NSD applications include multimedia information retrieval and our federated database application introduced in Chapter 1. In multimedia information retrieval the information being retrieved is not self describing. It is, for instance, not possible to arrive at a KB representation of a picture or sound without human intervention. We need someone to describe the picture or sound in natural language and we can then use this description to generate the KB representation. Similarly, for the locating and relating of information stored in federated databases it is not possible to use the names of schema objects to get a KB representation of information contained therein. This is made quite clear when one examines the types of names frequently given to schema objects, e.g. Table0051, Tax_SW_Tb, etc. Again an interactive session with a knowledgeable user is required to generate a KB representation of the information.

In the remainder of this Chapter the application of our system to both NSD and SD applications is further discussed. Section 6.1 describes the details involved in

applying our system to an NSD application and in Section 6.2 a general SD application is described. In Section 6.3 we make a decision on which application to use for evaluating our system. Conclusions on this Chapter are presented in Section 6.4.

6.1 NSD Applications

The discussion in this section concentrates on what have been termed as nonself-describing (NSD) applications of our approach to information matching. Multimedia information retrieval and the retrieval and relating of information in large scale federated databases have been cited as examples of NSD applications. There are many more NSD applications but for the purposes of the discussion here it is sufficient to focus our attention on just two of these applications. In order to further specify the multimedia application we can allude to a particular project being carried out at Dublin City University. The project involves an investigation into methods of performing multimedia information retrieval. In particular the participants are interested in the retrieval of images using caption descriptions. It is envisaged that this system could handle queries of the form : *'Show me pictures illustrating a landscape with trees'*. It is not difficult to see how this falls into the category of an NSD application of our system.

The intention in this Section is to describe a working model of how our system could be applied to an NSD application. To aid in this task we will discuss the operation of such a system in terms of what we refer to as the registration and querying procedures. The registration procedure is the name given to the generation of a KB representation of the information collection, for the FDBS application this is the information stored in component databases and for the multimedia application this could be a collection of images or pictures. We will refer to individual elements of the information collection as data sets. In any NSD application both the querying and registration procedures involve an interactive session with a user. For the registration

procedure the interactive session is with a knowledgeable user, perhaps the database administrator in the FDBS application or an art expert in the multimedia application. To register information the user simply enters a list of terms describing the information or data set being registered and similarly to query the system a user simply types in a list of terms roughly describing the requested information. The purpose of the interactive session is two-fold, firstly query and registration terms must be sense disambiguated, however, perhaps just as importantly, an interactive session allows the user to build up his query and registration terms dynamically. Our approach is based on the REFORM user-interface developed for the TINA text retrieval system, [Schw90]. In REFORM a terms context is determined using pre-determined head-modifier links in the text being processed. In contrast, we use the relational links in our WordNet derived KB for the same purpose. The querying and registration interface to an NSD application, expects a list of terms to be entered as input. The dialogue may be carried out at two levels, initially the correct HCGs must be determined, and following this, a context within these HCGs must be decided upon. A sample of a prototype registration/querying interface, is shown in below.

Please Enter a comma separated list of concepts : **Transport, Storage, Milk**

The concept '**Transport**' is ambiguous. Which sense is appropriate :

1. 'Act' as in 'something that people do or cause to happen ', As in :
 - Synset is { transportation shipping transport }
 - Glossary is ' the commercial enterprise of transporting goods and materials'
2. 'Psychological feature' as in 'a feature of the mental life of a living organism' . As in :
 - Synset is { ecstasy exaltation transport rapture }
 - IS A KIND OF -> { happiness gladness felicity }
3. 'Entity' as in 'something having concrete existence; living or non living '. As in :
 - Synset is { conveyance carrier transport }
 - Glossary is ' something that serves as a means of transportation '

Please choose appropriate senses : **1, 3**

The concept '**Storage**' is ambiguous. Which sense is appropriate :

1. 'Act' as in 'something that people do or cause to happen '. As in :

- Synset is { storage }
Glossary is ' the commercial enterprise of storing goods and materials '
- Synset is { storage }
Glossary is ' the act of storing something '

2. 'Entity' as in 'something having concrete existence; living or non living '. As in :

- Synset is { storehouse depot entrepot storage store warehouse }
IS A KIND OF -> { depository deposit repository }
- Synset is { memory storage store memory board }
Glossary is ' a memory and the CPU form the central part of a
computer to which peripherals are attached '

Please choose appropriate senses : 1, 2

The word form '**Storage**' as in 'something that people do or cause to happen' has 2 meanings :

1. Synset is { storage }
Glossary is ' the commercial enterprise of storing goods and materials '
2. Synset is { storage }
Glossary is ' the act of storing something '

Please Choose a Meaning : 1, 2

The word form '**Storage**' as in 'something with concrete existence, living or non-living' has 2 meanings :

1. Synset is { storehouse depot entrepot storage store warehouse }
IS A KIND OF -> { depository deposit repository }
2. Synset is { memory storage store memory board }
Glossary is ' a memory and the CPU form the central part of a
computer to which peripherals are attached '

Please Choose a Meaning : 1

The word form '**Milk**' has 3 meanings :

1. Synset is { milk }
Glossary is ' produced by mammary glands of female mammals for feeding their young '
2. Synset is { Milk Milk River }
Glossary is ' a tributary of the Missouri River '
3. Synset is { milk }
IS A KIND OF -> { dairy product }

Please Choose a Meaning : 3

Figure 6-1 A Sample Session with a prototype front-end

From the above user interaction, the correct context is determined and a small amount of term expansion takes place, (from an initial 3 terms to 15, if synonyms are included). Further expansion is possible by following the IS-A, HAS-PART, HAS-MEMBER, and HAS-SUBSTANCE links, (note the inverse of these links, IS-A-KIND-OF, IS-PART-OF, etc., along with the glossaries could be said to be used to define the context). Choosing terms for expansion could be decided by a combination of their information content values, their relative depths in the hierarchy, and the number of links emanating from them. A sample of how this expansion might operate would be :

The term '**Transport**' is quite broad in meaning, do you wish to specify [Y/N] : Y

There are two senses of '**Transport**' from which to expand :

1. Transport, Transportation, Shipping - as in the commercial enterprise of transporting goods and materials
2. Transport, Conveyance, Carrier - as in something that serves as a means of transportation

Please choose senses : 1, 2

Expanding '**Transport**' as in 'something that serves as a means of transportation'

1. Caravan - as in a group of wagons or pack animals travelling in single file
2. Dolly -- as in a wheeled platform for moving heavy objects
3. Litter -- as in a chair or bed carried on two poles by bearers
4. Mail -- as in a conveyance that transports mail
5. Public transport -- as in transporting passengers or mail or freight
6. Shipping, cargo ships, merchant marine, merchant vessels
7. Sidecar -- as in a small carrier attached to the side of a motorcycle
8. Ski tow, ski lift, lift -- as in carries skiers up a hill
9. Vehicle -- as in a conveyance that transports people or objects

Please choose from among the above kinds of 'Transport' : **5, 6, 9**

Expanding '**Transport**' as in 'the commercial enterprise of transporting goods and materials'

1. Hauling, trucking
2. Freight, freightage
3. Express, expressage - as in rapid transport
4. Moving - as in transportation of household or office belongings to a new address
5. Ferry, ferrying - as in transport by boat or aircraft

Please choose from among the above kinds of 'Transport' : **1, 2, 5**

Do you wish to specify further [Y/N] : **N**

Figure 6-2 A sample dialogue for expanding terms

This approach to querying and registration allows the option of weighting query and registration terms. It is, however, believed that this option should be discretionary. The weighting of terms, if correctly used, can be notably beneficial but in practice we believe many users would have difficulty generating appropriate weights.

6.2 SD Applications

Many of the applications introduced in Section 6 were self describing or SD applications. These applications are characterised by the fact the data sets making up the information collection are natural language text and as such it is possible to automatically generate a KB representation. In this Section we will discuss how our system could be applied to an SD application and in particular, how it differs from its employment in an NSD application. As with the discussion on NSD applications we will focus our attention on just two applications. For our discussion on SD applications we will concentrate on traditional document retrieval and information filtering. There are both information filtering and document retrieval projects currently being developed at Dublin City University. The document retrieval project basically involves the retrieval of Newspaper articles from the Wall Street Journal text corpus using a fixed set of queries, (from TREC, [Harm93]). The information filtering project involves the filtering of articles from a regularly updated on-line copy of the Irish Times, (an Irish national newspaper). The user prepares a short caption of natural language text referred to as a user profile which describes articles of interest to that particular user. The user profile is then compared against articles from the newspaper and articles found relevant to the user profile are filtered out for the user's attention.

The registration and querying procedures for the SD application just described are quite different to the NSD applications discussed in the previous Section. For both these applications the registration and querying procedures can be automated³⁵. Unfortunately, however, this automation is not necessarily a straightforward task. A mechanism is needed to read through the data set text or query text and isolate content

³⁵ In general, however, document retrieval systems can be queried interactively. For such systems, the prototype front-end discussed in the Section 6.1 is applicable.

describing terms that can be used to build the KB representations of either the data set or query. These terms must then be disambiguated, and as we have already seen WordNet makes very fine sense distinctions effectively making this a particularly difficult procedure.

In the following Section we will present and justify our choice of application for the evaluation of our system. Chapters seven and eight contain a more detailed discussion on this chosen application.

6.3 Choice of Evaluation Application

Before discussing our choice of evaluating application we will briefly discuss what we are evaluating. So far in this thesis we have developed a KB and two semantic similarity estimators. The purpose of the evaluation should thus be to determine whether :

- the KB is suitable for its intended use, as a controlled vocabulary in the representation of both information requests and requested information
- the semantic similarity estimators increase performance over a system using pattern matching as the comparison mechanism.

With this in mind we believe an SD application is better suited to the evaluation procedure. Reasons for this choice are centred around the following related points.

6.3.1 Availability of a test bed.

This point has quite simply to do with the fact we have no information collection for an NSD application. For the multimedia application we did not have access to a large enough test bed of multimedia objects to construct a multimedia database for our evaluation. For the FDBS application a large scale FDBS is required and although discussions took place with the Irish Department of Health with a view to using their databases as a test bed, the amount of effort involved in organising such a project was deemed too much. The developers of SSM, (refer to Section 2.3), also encountered this problem. They opted to simulate a large scale FDBS and to evaluate SSM in terms of costs and overheads of processing time and network communications. The obvious drawback of such an evaluation is the fact that the question of whether it works is not answered.

In contrast, Dublin City University is involved in the TREC project and as a result has an extract WSJ text corpus along with a set of TREC queries that could be used in a document retrieval SD evaluation application. Similarly the existence of the Irish Times filtering project provides a test bed for the use of an information filtering evaluation application.

6.3.2 Human Factor

In any evaluation of a computer system where a user interface is not being evaluated if there is a choice between a configuration involving human input and one that is completely automated, the automated configuration will in general be chosen. This is because human interaction introduces an unquantifiable error component in any experiment. The objectiveness and consistency of human decisions always make the results of such experiments questionable. In any NSD application humans are needed in both the registration and querying procedures. For the registration procedure a knowledgeable user is required. However in an SD application everything can be fully automated thus introducing a general consistency in the systems operation.

6.3.3 Evaluation Procedure

It is difficult to arrive at a procedure for evaluating the KB and similarity estimators using an NSD application. In [Chen92] a two phase experiment based on human recall and recognition tests is used to evaluate an information retrieval KB. However, the human factor argument of Section 6.3.2 is again pertinent here. In terms of evaluating the effectiveness of our similarity estimators, the only option available with an NSD evaluation application would seem to be further comparisons against human judgements of similarity. But as stated above, the semantic similarity estimators are to replace pattern matching as a comparison process and as such, our system should be compared against an information processing system using pattern matching and not human judgements of similarity. Unfortunately, all pattern matching information processing systems necessarily operate on SD applications.

In a document retrieval SD application it is possible to compare the performance of our semantic information processing system against a traditional pattern matching information retrieval system. The TREC project provides an automatic evaluation mechanism based on precision and recall, (refer to Chapter 9), which can be used to evaluate the performance of our system against a baseline pattern matching system.

6.3.4 Information Volume

This point relates back to the point made in Section 6.3.1. Although there is no threshold in terms of the volume of information needed for the evaluation, it is preferable to have a relatively large amount in order to support conclusions. For the document retrieval application we have half a gigabyte of text for the information collection and for the information filtering application over 300 megabytes of text was

available. It would be next to impossible to generate this sort of volume of information for an NSD application and as was stated in Section 6.3.1 no existing information collection was available.

Although all of these point are in favour of the use of an SD application, the one great disadvantage of such applications is the need for automatic sense disambiguation. We will see how this is handled in the next Chapter.

Within the self describing applications we singled out document retrieval over information filtering because of ease of evaluation. Evaluation of the information filtering application would necessarily have involved the use of humans, whereas the TREC evaluation mechanism both standardises and automates the evaluation process. So we conclude with the fact we are taking a non-interactive self describing document retrieval application as the evaluation task.

6.4 Conclusions

In this Chapter we enumerated a number of possible applications of our semantic information processing system. The main purpose of the Chapter was, however, to decide upon an application to evaluate our approach to information processing. Non-retrieval applications of our system were introduced in Section 6, however, we later categorised applications as being either self describing or non-self describing. These application categories and examples of them were further discussed in Sections 6.1 and 6.2. Finally in Section 6.3 we chose the self describing application of document retrieval to evaluate our system.

In the next Chapter there will be details of how articles from the WSJ text corpus and TREC corpus were processed to automatically generate registration terms.

Chapter 7 - Generation of KB Representations

7. Introduction

In the previous Chapter it was decided to use the Wall Street Journal and TREC queries in a document retrieval system to evaluate our semantic KB approach to information processing. Basically the system is given a TREC query as input and is required to rank WSJ articles with respect to their relevance to this query. Various configurations of the system, specifically the use of the information based versus the conceptual distance semantic similarity estimator, will be evaluated in this operation. However, the main purpose of this evaluation is to compare our approach against a traditional pattern matching information retrieval system.

The discussion in this Chapter concentrates on how we automated the process of converting WSJ articles and TREC queries into KB representations. Section 7.1 describes how the text of articles and queries were processed to remove as many non-content bearing terms as possible and prepared for input to the sense disambiguator. In Section 7.2 there is a description of our approach to sense disambiguation which includes a brief informal evaluation of the overall procedure. Conclusions are presented in Section 7.3.

7.1 Text Preprocessing

The text corpus used in our evaluation application is made up of 550 Mbytes of newspaper articles from issues of the Wall Street Journal (WSJ) between the years 1986 and 1992. The newspaper is published every weekday with around 380 documents per day. The articles themselves vary greatly in terms of subject, length, and writing styles. Refer to appendix F for further statistics of the text corpus. The articles are formatted into a pseudo-SGML structure which tags fields such as title of document (<HL>), Author (<AUTHOR>), place of writing (<DATELINE>), unique document identifier (<DOCNO>), the natural language text body (<TEXT>), etc.. All documents have a beginning and end tag (<DOC> and </DOC> respectively).

The 50 TREC queries are made up of natural language statements which specify more precisely what a user is looking for than the traditional approach of using a set of keywords. As with the WSJ articles each query is described using a number of different headings such as title, narrative description, concepts, definitions, etc.. An example query can also be found in appendix F. In this Section we tend to concentrate on the processing of WSJ articles, however, unless otherwise stated, all steps of the text preprocessor apply to both the text of WSJ articles and TREC queries.

Since the WSJ collection is so large, a significant amount of time is needed to solve the engineering problems associated with retrieving from such a large body of text and, as we shall see later, the procedures to perform sense disambiguation and query matching are very computationally expensive. The procedure to transform the raw text into a format suitable for input to the sense disambiguator should do as much as is possible to accommodate the sense disambiguation and subsequent query matching phases. With regard to the sense disambiguation procedure the text preprocessor should attempt to ensure there is a minimum of ambiguous terms and that proper nouns, acronyms or errors of the part of speech tagger should not find obscure meanings in the KB. The pre-processor should also endeavour to lighten the workload of the query matching phase by only keeping the most content bearing

document terms in the KB representation for the matching process. This Section describes the text pre-processor which was developed with these two goals in mind.

The raw text of articles and queries must go through a number of preprocessing steps before being semantically disambiguated. The order of these steps can be listed as follows :

- (1) Stripping of Headers and Trailers
- (2) Tagging of all words to indicate their syntactic category
- (3) Building up of collocations
- (4) Removal of non-nouns
- (5) Removal of nouns not occurring in the KB
- (6) Removal of non-content bearing nouns
- (7) Sorting of remaining index terms and removal of duplicates

Step one was straight forward and requires no further discussion. Step two was carried out using the RUCL syntactic parser and is discussed in Section 7.1.1. All other steps are discussed in Section 7.1.2.

7.1.1 Syntactic Analysis

The RUCL part of speech tagger is a domain independent syntactic parser developed as part of the SIMPR ESPRIT project, [Smar90]. What follows is a broad overview of how the tagger works, for a more detailed discussion refer to [Karl89, Vout92].

The RUCL tagger is made up of four components each of which performs a separate pass through the text in the process of syntactically tagging all words of the text. In the first pass the raw text is tagged with document, sentence and clausal linguistic boundaries. The output of this phase is input to the morphological

processor where words are decomposed into their base forms, suffixes, and prefixes. The morphological analysis allows for lexical ambiguity, so a term can be tagged with one or more readings or lexical interpretations. A lexical interpretation is made up of a word's part of speech, its inflections, and its syntactic function in the clause in which it occurs. The third pass involves context sensitive syntactic disambiguation of words with multiple morphological interpretations. This is carried out using the constraint grammar developed by Karlsson, which contains of the order of 1,100 disambiguation constraints, [Karl89]. Following this stage each word has one lexical tag unless it is truly lexically ambiguous. The final phase of the RUCL tagger appends syntactic functions for each word in a clause. These functions describe how a word affects and is affected by other words in the clause. Each function label is preceded by an "@" symbol. All words which modify other words are denoted with a "<" or a ">" symbol depending on the direction of the modification. For example the function label "@AN>" indicates that the current word is an adjective (A) and is modifying a noun(N) to its right(>). For the purpose of this present research and other research being carried out [Odon94], it was decided that these 32 function labels are too specific so these labels were grouped into six categories namely, heads, modifiers, verbs, adverbs, adjectives, and stopwords. The composition of these groupings can be seen in appendix E.

The version of the tagger used in our research was delivered as part of the SIMPR project and on average it takes approximately 4 hours to parse 1 Mbyte of natural language text from the Wall Street Journal, (using a SparcStation 2). A commercial version has been developed and marketed by RUCL which is reported to be many times faster. In tests carried out by RUCL on the part of speech tagger, it labelled only 3-6% of all words with lexically ambiguous tags and 99.7-100% of all words retain appropriate morphological readings. However, other evaluations of the RUCL tagger have not produced as good results, [Smea92b]. In our experience we found the RUCL tagger multiply tagged words on the vaguest presence of ambiguity. In a test we carried out involving 60 megabytes of text from the WSJ, (8 million words), we found 1.5 million words were multiply tagged, (19% of all words). The situation is made worse by the fact there is no ranking or scoring for these alternate

interpretations. The tagger also tends to make a number of outright mistakes. For instance, it is not unusual for the tagger to tag an obviously unambiguous word such as “sell” or “say” as a noun as opposed to a verb. A final criticism of the RUCL tagger is apparent in its tokenising of words. If the parsed text was displayed alphabetically sorted with one word per line, the following scenario is quite probable

```
monastery
money
money,
money.
mongrel
```

As can be seen the punctuation characters are included as the last characters of the word. This is bound to cause problems in any matching process that uses the text of documents and queries to attain index and index terms. Fortunately, however, many of these ‘tagger errors’ are trapped by subsequent steps of the text pre-processor.

7.1.2 Processing of Syntactically Parsed text

The remaining steps in the text preprocessing take the syntactically tagged text and further process it to prepare it for use by the semantic tagger and the retrieval engine, refer to Section 7.1.1 for a list of the steps.

The third preprocessing step involves building up KB collocations. The output of the tagger retains the sequence of the text so its possible to pass through this output attempting to locate co-occurring words that appear as a collocation in the KB. It was necessary to carry out this step before step (5) since many collocations contain non-nouns, (e.g. ‘department_of_defence’, ‘by_and_by’, etc.). Details of the

collocation detector can be found in Section 4.2.2. Building up collocations is a very good way of performing sense disambiguation. If, for example, the nouns 'transport system' appear in the text then both words are ambiguous, however, by combining both together in the collocation 'transport_system', the ambiguity is removed. Furthermore, it can be ensured that these collocations have an information content score because the entire WSJ was processed to include collocations in the calculation of information content scores, refer to Section 4.2.

The fourth and fifth steps of the pre-processor simply involve the stripping of all words not occurring in the KB. Many of the tagging errors are eliminated at these stages. It was found that a very large percentage of words tagged as nouns are found in the KB. This is another indication of the extent and exhaustiveness of WordNet. The presence of collocations did, however cause a problem with hyphenated words. Terms such as 'atom-bomb' and 'articulated-truck' appear in the KB as collocations and in order to deal with this it was necessary to replace hyphens with spaces whenever a hyphenated word was not found in the KB. The order of execution of steps four and five was reversed for the document titles. This was in response to the fact that many content bearing nouns were being incorrectly tagged as adjectives by the RUCL tagger. Although, nothing could be done about this situation as a whole it was decided that given the possible importance and the small size of the document titles the pre-processor should simply locate the title words in the KB regardless of their part of speech.

The sixth step of the pre-processor was concerned with increasing the speed with which the matching of queries and documents could take place. The simplest matching process would simply involve the pairwise matching of all query terms against all document index terms using some semantic similarity estimator. So, obviously, the fewer query and index terms there are the faster the procedure. Of course, the trick here is to keep the content bearing terms and remove only those terms that are superfluous to the essential nature of the document or query. Following a number of experiments it was decided to remove terms with any of the following characteristics :

- Having two or less characters
- Terms appearing in general Stop Lists
- Terms which are proper nouns but could have acronym or slang interpretations
- Terms with large document frequencies

Terms with two or less characters are very often either parser errors which had slipped through previous pre-processor stages or acronyms that may unexpectedly have found a match in the KB. Single character words such as 's', (possibly from an apostrophe s), or 'a' were often tagged as nouns and were subsequently found in the KB with meanings such as :

4 senses of s

Sense 1

{ sulfur, S, sulphur, atomic number 16 }
IS-A => { chemical element, element }

Sense 2

{ south, due south, S } : (the cardinal compass point that is at 180 degrees)
IS-A => { cardinal compass point }

Sense 3

{ schilling, S, Sch } : (Austria)
IS-A => monetary unit

Sense 4

{ mho, siemens, reciprocal ohm, S }
IS-A => { conductance unit }

4 senses of a

Sense 1

{ vitamin A, antiophthalmic factor, axerophthol, A }
IS-A => { fat-soluble vitamin }

Sense 2

{ angstrom, angstrom unit, A }
IS-A => { metric linear unit }

Sense 3

{ ampere, amp, A }
IS-A => { current unit }

Sense 4

{ A } : (the blood group whose red cells carry the A antigen)
IS-A => { blood group, blood type }

Similarly, two lettered acronyms were often found to appear in an unlikely guise in the KB. One particularly commonly occurring example of this was the word 'co', which, given the general domain of the WSJ, would be assumed to stand for 'company', however, the meanings offered by the KB were :

2 senses of co

Sense 1

{ carbon dioxide, CO, carbonic acid gas } : (a heavy odourless gas)

IS-A => { dioxide } : (an oxide containing two atoms of oxygen)

Sense 2

{ cobalt, Co, atomic number 27 } : (a ferromagnetic metal)

IS-A => { chemical element, element }

The second category of 'non-content bearing terms' were those appearing in previously compiled lists of non-content bearing words. Fox reports on an exercise to generate a stop list of non-content bearing words in [Fox90]. Many of these words are frequently occurring words in text or words that relate to points in time. Words such as 'today', 'tomorrow', and 'yesterday' are all relative terms and provide little information when it comes to determining relevance to queries. Besides which, one visible characteristic of TREC queries is the absence of a temporal dimension. Not all of the words in Fox's stop list were useful since many of them were non-nouns, however, any of them that occurred in our KB and were not already included in our stop list were added to our exception list of index terms, see appendix G for a list of these terms.

The third class of terms to be included in those exempt from use as terms in the KB representations of queries and articles are characterised as being proper nouns with a slang or acronym interpretation. Although the RUCL parser had a special tag for proper nouns, we decided not to use it. In hindsight this is probably just as well since place names, which are particularly valuable in query matching and widely covered in WordNet, would be classed as proper nouns. As such, the option of simply excluding all proper nouns would result in the loss of a lot of valuable information.

Following examination of the words that fall into the set to be excluded it was found that most are names of people as opposed to place names. If we take, for example, the name 'John', it would be reasonable to expect an occurrence of this word in the text to be excluded from the set of representative terms following step five of the preprocessor, however, the meaning :

{ toilet, lavatory, can, head, facility, john, privy, bathroom } :
IS-A => { room }

ensures it remains. Other examples include James, Ken, and IRA (intended meaning of Inland Revenue Association) :

3 senses of james

Sense 1

{ James, William James } : (1842-1910)

IS-A => psychologist

IS-A => philosopher

Sense 2

{ James, Henry James } : (1843-1916)

*IS-A => { writer, author } : (writes (books or stories or articles or the like)
professionally (for pay))*

Sense 3

{ James, James River } : (a tributary of the Missouri River)

IS-A => { river } : (a large stream of water)

1 sense of ken

Sense 1

{ cognizance, ken } : (range of what one can know or understand)

IS-A => { knowing } : (clear and certain mental apprehension)

1 sense of ira

Sense 1

{ wrath, anger, ire, ira }

IS-A => { mortal sin, deadly sin }

Given the size of WordNet there are many more of these types of misinterpretations possible.

The final category of exception terms is constructed by making use of a well known information retrieval technique for determining how content bearing a particular term is. The inter-document frequency of a term is a count of the number of documents in the document collection the term appears in. It is commonly held that terms with a high inter-document value are not particularly content bearing. This is fairly intuitive if one considers that terms with a high inter-document value are not particularly good at discriminating the content of individual documents from each other. This low discriminatory ability makes these terms poor index terms in the query-document matching process. Previous studies in this area had set the threshold under which index terms could be included at values between ten and thirty percent. This meant that any term occurring in more than this threshold percentage of the documents would not be included as an index term. A number of experiments were carried out to determine the optimum threshold for our purposes. Two WSJ files, each about one and a half megabytes, were used for the test purposes. A threshold of 30% was tried initially. With over 154,000 documents in the WSJ collection a threshold of 30% means that a term occurring in 46,200 or more documents was regarded as non-content bearing. At this threshold only six terms were found for both test files; 'make', 'million', 'month', 'much', 'share', and 'take'. Given the test collection, it would seem reasonable to leave these terms out as index terms. At a threshold of 10% there were 61 unique terms that qualified as appearing in more than 15,400 documents, (see appendix H). Again the same 61 terms were returned for both test files and each of the terms would be regarded as non-content bearing given the general domain of the collection. At 5% the number of exception terms more than doubled and the set of terms above the threshold for both files differed very slightly. The extra terms due to the lowering of the threshold aren't all as clearly non-content bearing as those at the 10% threshold. The final threshold was thus set at 10%.

Apart from excluding terms with two characters or less each of the other approaches to reducing the number of terms in the KB representation were collectively used to construct a large exception list of terms to be excluded. In order to give a feel for the actual degree to which this reduced the number of terms for both the sense disambiguation and query matching phases, a few statistics were gathered on

the WSJ collection. Of particular interest was the statistic showing the average number of terms per document. A term here refers to all non-stop list words, (where the stop-list is one from a typical stemmer such as Porter's, i.e. words such as 'a', 'the' and conjunctions). The WSJ collection is divided into 710 files of approximately one megabyte each, (although this can vary from between 0.7 of a megabyte to 1.6 megabytes). Statistics were gathered for blocks of 50 files. The following bar chart illustrates the average number of terms per document for each fifty file block, (further statistics on the WSJ collection can be found in appendix F).

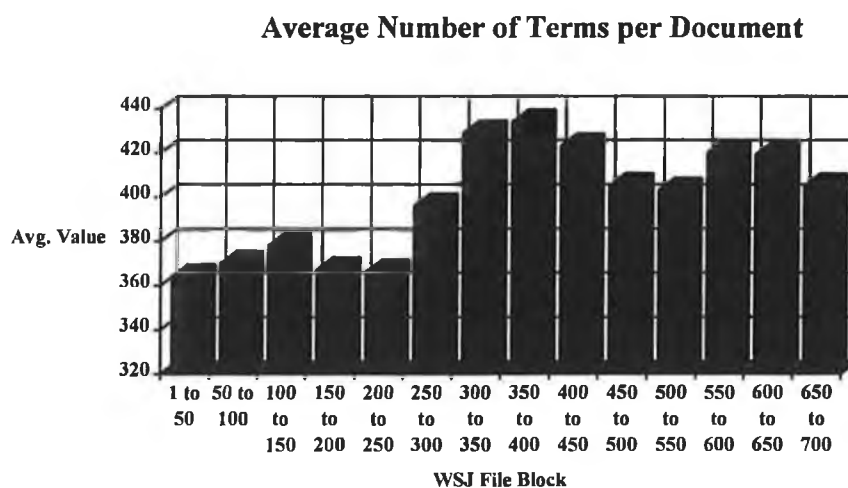


Figure 7-1 Average Lengths of WSJ articles

As can be seen from the chart the average number of terms per document ranged between 365 and 435. The actual range between the article with the largest number of terms and that with the smallest number of terms was, however, very big with the maximum number of terms at around 8,500 and the minimum number of terms at 5 or 6, (these small articles are literally one liners which were possibly accompanied by a picture). Two WSJ files, each with around 400 WSJ articles, were then pre-processed³⁶. The resulting average number of terms per article was reduced to 135 and 119 respectively. This accounted for a considerable reduction of article terms qualifying as index terms in the KB representation of an article. The time taken to

³⁶ It took about ten minutes to preprocess a WSJ file

carry out the sense disambiguation and query matching was thus considerably reduced.

The final step of the preprocessor involves the doubling up of duplicate index and query terms. It is important to note this step occurs following sense disambiguation and as such is only a preprocessing of the text for the retrieval engine. It is obvious that if a particular term is repeated several times throughout the text it is only necessary to compute a similarity score for it once and then to multiply this value by the number of occurrences. The repetition of index terms in an article is quite common, this is perhaps not surprising given the fact articles tend to discuss specific topics using the vocabulary of that topic. On average, this process reduced the number of index terms by approximately 10%.

7.2 Sense Disambiguation

The discussion in this Section is centred on the process of semantically tagging the document index terms with the appropriate KB synsets. Section 7.2.1 gives a brief review of what little work is being done on sense disambiguation using WordNet. Although much work has been carried out on word sense disambiguation in general, we restrict our discussion to that of word sense disambiguation using WordNet. Section 7.2.2 goes on to describe the methodology developed for this research. A small example containing three test articles is presented in this Section. Again, the discussion concentrates on the semantic tagging of WSJ articles, however, the procedure for queries is the same except on a far smaller scale. Finally Section 7.2.3 suggests possible improvements to the approach and discusses some of its limitations.

7.2.1 Sense Disambiguation in Information Retrieval

Recently there has been a heightened interest in the application of word sense disambiguation to information retrieval. Many researchers held the opinion that the precision of retrieval results could be greatly improved if the false matches, caused by the ambiguity of polysemous words, were eliminated through word sense disambiguation. Krovetz and Croft, [Krov92], were perhaps the first to do major research into the effect of word sense ambiguity document retrieval systems. They found that a sense mismatch was more likely to occur when the document was non-relevant and furthermore, sense mismatches were more likely to occur when there are a small number of words in common between the query and document. Thus they concluded that the impact of sense ambiguity on information retrieval was not dramatic, but disambiguating word senses was probably beneficial to retrieval when there were index and query terms in common. The relationship between ambiguity and retrieval performance was further investigated in a set of experiments reported in [Sand94]. In these experiments ambiguity was artificially introduced to the text of the documents using pseudo-words, [Yaro93]. The procedure involved replacing all occurrences of randomly selected terms with their collocation. Retrieval was then performed on both the original collection and on the collection with the pseudo-words. Surprisingly, there was no great degradation in performance due to the introduction of ambiguity. Further experiments were carried out to see the effect of performing sense disambiguation on retrieval performance. Since the ambiguity was artificially introduced it was possible to disambiguate to a controlled degree of accuracy. It was found that at 75% accuracy the results were considerably worse than those for the fully ambiguous collection. Only at 90% accuracy did the results approach those of the ambiguous (and original) collection. It was thus concluded :

‘Traditional IR systems are relatively insensitive to ambiguity but very sensitive to erroneous disambiguation. ... tools built for computational linguistics tasks need to operate at, at least 90% accuracy before they are of practical use’.

This would help explain the bad results reported in [Voor94]. Voorhees also found that missing correct matches because of incorrect sense resolution has a much worse effect on performance than making false matches.

It would thus seem that word sense ambiguity is not as important an issue in the performance of information retrieval systems as would have been intuitively considered. This certainly is the case with traditional or standard information retrieval systems, however, for the knowledge based semantic information retrieval proposed in this research, sense disambiguation is a vital component. Since we are using concepts from the knowledge base to represent documents and queries it is essential to know which concepts to use. The very nature of our mechanisms to estimate the similarity of terms further require that exact concept senses are used and not any of a number of different concept nodes. Furthermore, due to the fine sense distinctions present in our KB the issue of word sense ambiguity is more the rule than the exception. Unlike a traditional IR system where the ignoring of ambiguous words has little effect on retrieval results, in a system such as ours, such a course of action would effectively eliminate a large percentage of the terms used to represent articles and queries. In experiments involving ten batches of 1000 WSJ articles it was found that on average 75% of the terms had more than one sense in our KB and on average each term had 3.1 senses.

In terms of sense disambiguators developed using WordNet, only Sussna, [Suss93], and Voorhees, [Voor93], are described in any great detail. Sussna, in [Suss93], attempted to automatically sense disambiguate articles from the Time Magazine on-line collection. His approach was to use the surrounding words as a context and to choose the sense of the ambiguous word that was on average closer to all the words in the context. Closeness between words was estimated using a conceptual distance similarity estimator similar to the one used in our research. Results from his approach were, in general, quite good, with an optimum window size found to be 41 words. One criticism, however, of his approach is the computational overhead. The approach would not be suited to an application involving large quantities of text, (as we have in our application). Voorhees' approach is also based

on the notion of locality. She tries to categorise the senses of words, the sense belonging to the category of the majority of context words being chosen as the intended sense. Word categories are defined in WordNet using constructs referred to as hoods. A hood of a given synset, *s*, is the largest connected subgraph that contains *s*, contains only descendants of an ancestor of *s* and contains no synset that has a descendant that includes another instance of a member of *s* as a member. A hood is represented by the synset that is the root of the hood, (from Figure 7-2):

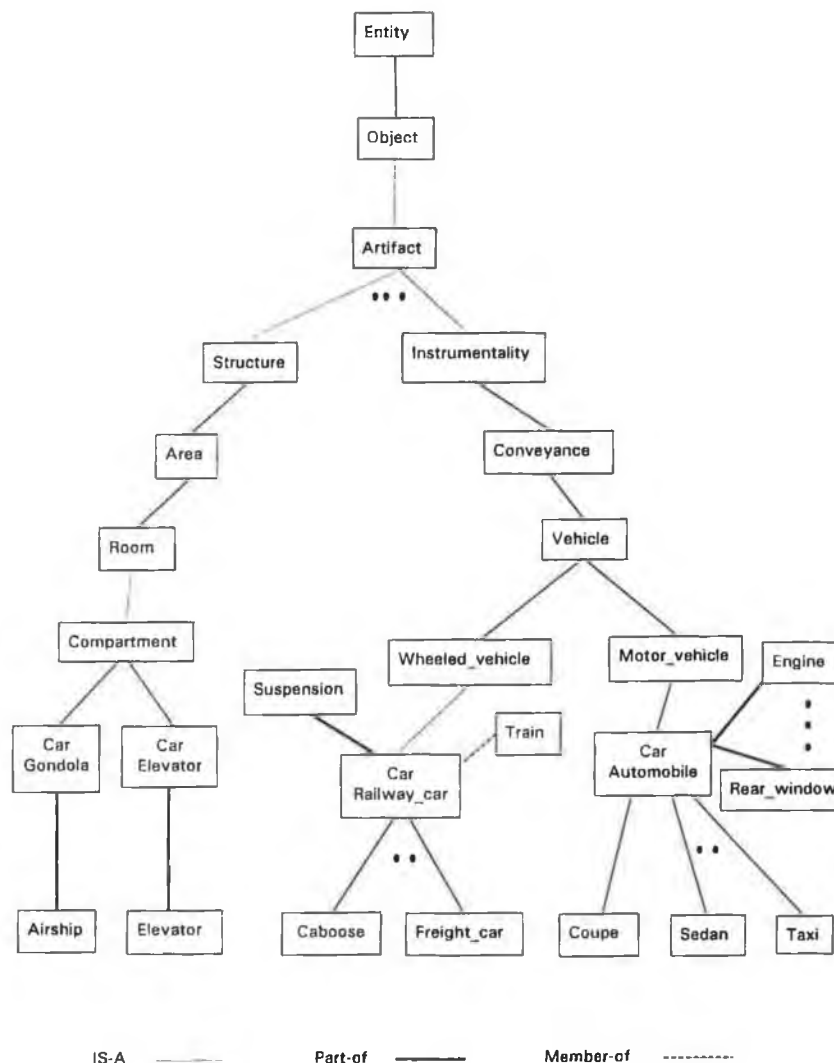


Figure 7-2 WordNet Extract for the concept 'car'

the hood for the railway sense of car is wheeled vehicle, the hood for the automobile sense of car is motor vehicle, and the hoods for the gondola and elevator senses of car are themselves.

There was no systematic evaluation of Voorhees's disambiguator because of the sheer amount of work involved in such a procedure. However, from a subjective evaluation she says :

‘ ... the technique is not a reliable method for choosing among the fine sense distinctions WordNet makes’, [Voor93].

One criticism of the approach is the fact that no preference is given to the sense that occurs most frequently in a text corpus. As will be seen in the following Section, this is not the case in our approach. However, perhaps the major drawback of her approach, (and of Sussna), is the fact only one sense is chosen as the intended sense. As Voorhees points out, sense distinctions in WordNet tend to be subtle in many cases and any of a number of senses could be seen as being appropriate. This point is discussed further in the following section.

7.2.2 Method

The important point about the approach to sense disambiguation taken in this research is that its not always assumed there is only one correct sense. Instead it is believed that in many situations a number of senses may be appropriate, [Sutc91]. This is particularly true in WordNet's case where multiple senses of a word are present simply because the word can appear in different points in the network and not because there are a number of different semantic meanings, Figure 7-3 illustrates.

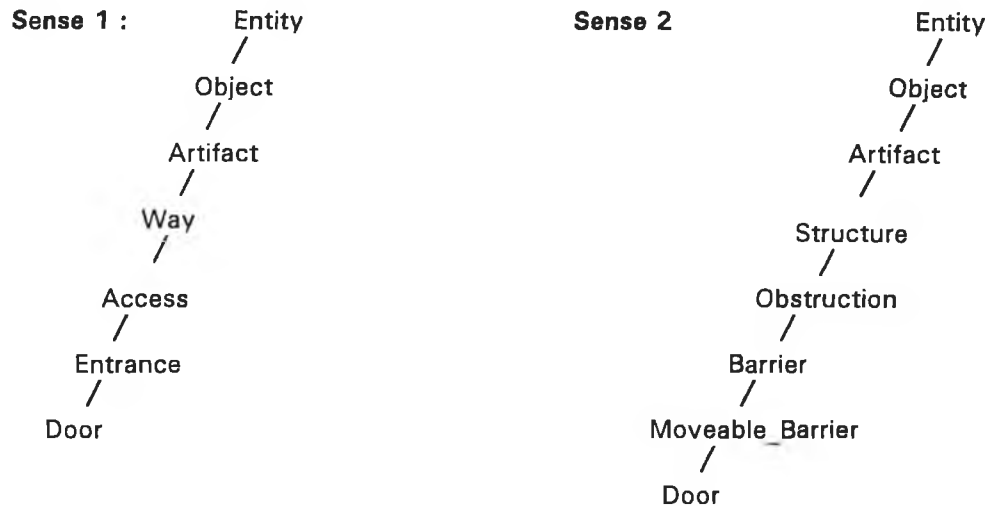


Figure 7-3 Entries for 'door'

What is suggested for the sense disambiguation here is that ambiguous words are tagged with all relevant senses with each sense getting a relevance score. Relevance scores are determined using the following sense disambiguation procedures :

- (1) Synonyms
- (2) Glossary
- (3) Locality
- (4) Information Content

Context is captured by using the surrounding nouns. A moving window into the text is effectively created and the middle term in this window is disambiguated. The size of the window is constant, (currently set at 11), since following the stripping of words not appearing in the KB it can always be ensured that all words being disambiguated will have the same number of surrounding nouns as a context. The first two disambiguation techniques work by simply looking up each of the other context words in the synonym and glossary listings of the various senses of the word being disambiguated. This is somewhat reminiscent of Lesk's method of sense disambiguation, [Lesk86]. The technique is based on the assumption that if a context

concept is found in the glossary entry of a particular sense of the ambiguous concept, then this sense is likely to be the correct one. The procedure can be made clearer with the following example. If we take 'Terminal' as the ambiguous concept, and 'Computer' and 'Workstation' as the context concepts, we have the following KB ambiguous senses of terminal :

1. Synset { Terminal }
Glossary : 'An attachment on a wire or battery or other electrical device for convenience in making a connection'
2. Synset { Terminal }
Glossary : 'An input-output device providing access to a computer: has a keyboard and display'
3. Synset { Terminal Terminus Depot }
Glossary : 'Where transport vehicles load or unload passengers or goods'

Now, it can be deduced that the second sense is the correct sense given the fact the context concept 'Computer' appears in its glossary. This mechanism could prove to be quite efficient given the fact that all concepts in a context tend to describe the same piece of information. However, in practice it has been found that the likelihood of either the glossary or synonym disambiguators being successful is unlikely. If either disambiguator is successful in finding a match, the particular sense in question should get a very high weighting. The size of the weighting will be affected by whether or not the matching context word is itself ambiguous or not.

The locality disambiguator amounts to something similar to Sussna's mutual constraint technique, [Suss93], although not exactly the same. It works as follows, given three terms t_1 , t_2 , and t_3 , with 3, 2, and 1 senses respectively and assume we are disambiguating term t_2 . We proceed by calculating the semantic similarity between each of t_2 's senses and all other terms in the context, (call this value DTOT). For ambiguous terms in the context we take the sense which is closest to the particular sense of t_2 for which DTOT is being calculated. Each of t_2 's senses is then given a relevance score according to its DTOT value with the lowest DTOT value getting the highest relevance score. The window size chosen for the locality disambiguator is of size 11. Obviously the larger the window size the slower the resulting process and unfortunately time is a primary factor in the disambiguating process. However, although the window size is only a quarter of the size found to be optimal by Sussna,

we believe that following the preprocessing of the text, as was discussed in Section 7.1, a window size of 11 in this 'concentrated text' is equivalent to a far larger window in normal, unprocessed text. For reasons of time also, the semantic similarity estimator used is the information based approach. This approach is found to operate considerably faster than the conceptual distance approach.

The final disambiguator is referred to as the information content disambiguator. This method of disambiguation will probably work well for situations where one particular sense is far more widely used than another. Although both senses will get the same frequency value in the information content calculation the fact that one is more widely used than another would suggest it has more children nodes than the other. And since the information content of a node is derived partly from the information content values of its subordinate nodes means it would have a greater information content value. For example :

1. There are two senses of 'hill' in the KB :

- { hill } as in :

IS A KIND OF { natural_elevation elevation }

Information content score : 3.97445

- { mound hill pitcher's_mound } as in the glossary :

'(in baseball) the slight elevation on which the pitcher stands'

Information content score : 4.024407

2. There are a few senses of 'pen' in WordNet, two of which are :

- { pen } as in the glossary :

'a writing implement with a point from which ink flows'

Information content score : 5.088991

- { pen } as in the glossary

'a female swan'

Information content score : 5.234305

3. There are a few senses of 'transport' in WordNet, three of which are :

- { conveyance carrier transport } as in the glossary :

'something that serves as a means of transportation'

Information content score : 2.49417

- { transportation shipping transport } as in the glossary

'the commercial enterprise of transporting goods and materials'

Information content score : 3.492429

- { ecstasy exaltation transport rapture } as in :

IS A KIND OF { happiness gladness felicity }

Information content score : 4.389451

In all of the above cases the first sense is more likely than the others. This is reflected in their information content values, (note : the broader the term the lower the information content score). However, the difference in scores is generally quite small, effectively ruling out the direct use of information content scores in assigning relevance scores. Relevance scores will instead be assigned as pre-set weights with the sense with the lowest valued information content value receiving the highest weight, and so on.

The use of the information content scores as a disambiguation technique is particularly apt if the same text corpus was used to calculate the information content values. Take for example the concept 'bank', with the ambiguous meanings river bank and commercial bank. It would be expected that the sense commercial bank would have a better information content value than river bank if the text corpus was inclined toward the financial domain.

The information content disambiguator is, however, not without fault. Due to the fact that 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 polysemous words. This is illustrated in the following example, (previously introduced in Section 5.2.2); the concept '*fruit*' has three meanings in our KB :

```

Sense 1      (3.374811)
{ fruit } -- (the ripened reproductive body of a seed plant)
=> { reproductive structure }

Sense 2      (4.600377)
{ fruit } -- (edible part of a seed plant esp. one having sweet flesh)
=> { produce, green goods }

Sense 3      (3.217927)
{ yield, fruit }
=> { product, production }.

```

When one thinks of the concept *fruit* it is more than likely in terms of either sense 1 or 2 above, (as opposed to sense 3). However, the information content sense disambiguator chooses the third sense as the intended sense, (the information content values of each sense is in brackets). If we examine each sense in the KB we discover that senses 1 and 2 have many children nodes but sense 3 is a leaf node. How then

does sense 3 have a broader information content value than the first two senses ? The answer can be found in the fact the information content value of the third sense of fruit is calculated both from the frequencies of the word '*fruit*' and the word '*yield*', in a large text corpus. As can be imagined, the word '*yield*' occurs quite frequently in the WSJ corpus and, as a result the synset { *yield*, *fruit* } gets a broad information content score. To accommodate this finding we altered the information content disambiguator to take into account situations similar to the one just described³⁷. Basically, if a synset contained a number of words an attempt was made to discover the predominant word. Predominance was determined using the familiarity fields of WordNet, (refer to Section 3.1.2 for a discussion on WordNet fields).

The four disambiguation techniques are combined to give an overall weighting for the individual senses. Through informal experimentation and intuitive reasoning it was decided to weight each disambiguator as follows. If an unambiguous window context word is found in the synonym list of one of the senses of the word being disambiguated, that sense is automatically given a score of 10. If the context word is itself ambiguous, the sense in question is given a score of 3. The respective scores for the glossary disambiguator are 7 and 2, reflecting the intuitively weaker power of this disambiguator. With respect to the locality and information content disambiguators it was decided to weight the top five senses of an ambiguous word. It was also decided that both disambiguators were of roughly the same disambiguating power and, as such, should have the same weighting values. The values assigned to the top five senses were 5, 3, 1, 0.5, 0.25. If, however, the information content disambiguator is working with a synset which contains multiple words and the concept being disambiguated is not the predominant word of the synset then, depending on the relevance of the sense being operated on, it is assigned one of the following five scores, 4, 2, 0.5, 0.25, 0.125.

³⁷ It should be noted that on average there is only one word per synset, so situations similar to the one described don't occur that often

It took from between 18 and 20 hours of CPU processing time to disambiguate one thousand WSJ articles, (approximately 250,000 noun occurrences). The output of the disambiguator was a file of the following format :

WSJ_word Num_of_occ Num_KB_senses { KB_synset HCG_file Ambig_weight }.

The list of senses for an ambiguous WSJ word are sorted in order of the final weights assigned to them by the sense disambiguation process. As such the ‘*winning*’ sense is first in the list and the next most likely sense is next, and so on.

7.2.3 Sample Text

In this Section there will be a brief example of the sense disambiguator operating on a sample text collection. The sample was actually designed to test the retrieval engine, (refer to Section 8.2), and as such doesn’t highlight all nuances of the sense disambiguator. However, a larger example would be difficult to present here and, besides which, the more important aspects of the disambiguation process are successfully displayed. The sample collection is made up of the following three articles :

ARTICLE : Fruit Article	ARTICLE : Cutlery Article	ARTICLE : Transport Article
\$ST	\$ST	\$ST
fruit *	\$ET	vehicle
\$ET	cutlery *	transport *
banana *	fork *	\$ET
food *	knife	car *
apple	spoon *	sedan *
pineapple	soup spoon	truck *
\$ED	\$ED	delivery van
		cart *
		articulated lorry
		bicycle
		\$ED

Table 7-1 Test Articles

The above format is, in fact, the exact output format of the text pre-processor for WSJ articles. Every article begins with the keyword 'ARTICLE' followed by the article identifier, typically the date the WSJ article was written followed by the story number. Following the article identifier is the begin article identifier, \$ST, this is followed by the title terms which are, in turn followed by the end of title identifier, \$ET. As can be seen from above, the 'Cutlery' article contains no title, this is not unusual among WSJ articles. Index terms from the main body of the article follow the end of title identifier and the article is ended by the end of article identifier, \$ED.

Of particular interest in this Section is what senses the disambiguator chose as the correct sense. Ambiguous terms are marked by an asterisk. As can be seen there is a particularly high percentage of ambiguous words, and many words that would be thought of as being unambiguous are ambiguous. This is mainly due to the large number of commonly used words that have a single, general meaning, but may have a large number of specific interpretations within this general meaning. Due to WordNet's exhaustiveness, all these interpretations are included as separate senses. This is illustrated in figure 7.2 with the word 'door' and can be seen in the sample articles with words such as 'food', 'fruit', and 'banana'. The text below presents the options available to the semantic tagger for the ambiguous words in the first article :

3 senses of fruit

Sense 1 [10.00]
{ fruit } -- (the ripened reproductive body of a seed plant)
=> { reproductive structure }

Sense 3 [7.00]
{ yield, fruit }
=> { product, production }

Sense 2 [2.00]
{ fruit } -- (edible part of a seed plant esp. one having sweet flesh)
=> { produce, green goods }

2 senses of banana

Sense 1 [12.00]
{ banana } -- (elongated crescent-shaped yellow fruit with soft sweet flesh)
=> { fruit } -- (edible part of a seed plant esp. one having sweet flesh)

Sense 2 [8.00]
{ banana, Musa sapientum }
=> { monocot, monocotyledon, liliopsid } -- (flowering plant)

2 senses of food

Sense 1 [10.00]

{ food, nutrient } -- (any substance that can be metabolized by an organism to give energy and build tissue)
=> { substance, matter } -- (the tangible stuff of which an object consists)

Sense 1 [6.00]

{ food, comestible, comestibles, edible, edibles, pabulum } -- (any substance that can be metabolized by a living organism into energy and body tissue)
=> { substance, matter } -- (the tangible stuff of which an object consists)

The numbers in bold after each sense gives the scores given to that sense by the disambiguator. As can be seen the correct choices were made in all three situations, however, with regard to the words 'fruit', and 'food', more than one sense could be deemed as being appropriate. The disambiguation of the word 'fruit' was discussed in the previous Section. In its context here, the locality disambiguator chose sense number 1, and as expected the information content disambiguator chose the second sense. There was no tied ranking, however, because the word fruit was not the dominant word in the { yield fruit } synset³⁸. The results of disambiguating the second article are as follows :

2 senses of cutlery

Sense 1 [8.00]

{ cutlery } -- (implements for cutting and eating food)
=> { tableware } -- (utensils for use at the table)

Sense 2 [8.00]

{ edge tool, cutlery, cutting tool }
=> { cutting implement } -- (a tool used for cutting or slicing)

4 senses of fork

Sense 1 [8.00]

{ fork, tablefork }
=> { cutlery } -- (implements for cutting and eating food)

Sense 2 [7.00]

{ branching, ramification, fork, forking }
=> { division }

Sense 3 [1.50]

{ furcation, bifurcation, fork }
=> { angular shape } -- (a shape having one or more sharp angles)

Sense 4 [0.75]

{ crotch, fork } -- (the angle formed by the inner sides of the legs where they join the human trunk)
=> { angle }

³⁸ Interestingly, 'yield' wasn't the dominant word either. Unless there is a considerable difference between the familiarity values of words in a synset then no one word may be deemed as being predominant

2 senses of knife

Sense 1

[10.00]

{ knife } -- (a cutting instrument with a sharp-edged pointed metal blade with a handle)
=> { edge tool, cutlery, cutting tool }

Sense 2

[6.00]

{ knife } -- (weapon with a handle and blade with a sharp point)
=> { weapon, arm, weapon system }

3 senses of spoon

Sense 1

[12.00]

{ spoon } -- (a piece of cutlery with a shallow bowl and a handle; used to stir or serve or take up food)
=> { cutlery } -- (implements for cutting and eating food)

Sense 2

[6.00]

{ spoon, spoonful }
=> { containerful }

Sense 3

[4.00]

{ spoon } -- (a golfing wood with an elevated face)
=> { wood } -- (originally a golf club with a wooden head; metal woods are now available)

The results here are a little more interesting. The two meanings of ‘cutlery’ are not that far apart, and again either interpretation may be seen as valid. As can be seen the sense disambiguator gives them both the same score. What actually happened here is the information content disambiguator chose one sense as the context independent ‘winner’ and the locality disambiguator chose the other, (refer to Section 8.2 to see how these tied ‘winners’ are handled by the retrieval engine). Cutlery was found to be the dominant word in both synsets so both disambiguators had the same power. There is a certain difference in meaning between both KB senses of knife, one is intended as a weapon and the other as a cutting tool, and the disambiguator made the correct choice here. Looking at the word fork, we see that the same situation has arisen as with the term ‘fruit’ in the first article. The locality disambiguator chose the first sense as the intended sense, whereas the information content disambiguator chose the second sense. However, as with the term ‘fruit, the word ‘fork’ is not the dominant word of the synset { branching ramification fork forking }, and as such, the less powerful weighting set is used by the information content disambiguator. Finally, for this article it can be seen that all three senses of ‘spoon’ are quite different, and again the disambiguator made the correct choices. The results for the final article are as follows :

Chapter 4 - KB Construction

4. Introduction

In Chapter 3 we described the lexical database WordNet. In this Chapter we describe how WordNet is extended to become the KB in our semantic information processing system. The extensions involved are firstly, the addition of a field to approximate the information content value of a synset and secondly, the weighting of the relational links between nodes. To facilitate these extensions as well as making WordNet more conducive to our application we modified its organisation from a single massive semantic network into a number of hierarchical concept graphs, (HCGs).

Details of how the HCGs were constructed can be found in Section 4.1. The discussion gives a complete overview of the KB with a description of the HCG construction process, the organisation of information within HCGs, and the indexing method used to access this information. In Section 4.2 there is an in depth discussion on how we arrived at an approximation for the information content of a synset. Included in this discussion are details of how our KB of HCGs was extended to include an information content field. Finally Section 4.3 presents our automatic link weighting mechanism.

4.1 Building of HCGs

WordNet as described in the previous Chapter is a single massive semantic network. In this Section we discuss how we split WordNet into a number of separate

concept graphs. Reasons for doing this are centred around a need to increase speed of access and general manageability problems posed by trying to make extensions to WordNet as a single unit. However, other advantages to partitioning WordNet will become apparent in Chapters 6, 7, and 8. One of the problems of WordNet's simple disk based organisation is the amount of CPU time it takes to access information. In any large scale application using WordNet where there are many thousands of accesses to concept nodes in the network, such access delays cause operational bottlenecks. It is not difficult to see how by splitting the network into a number of component parts the speed of access to concept information can be greatly increased. Also, the proposed extensions to WordNet described in Sections 4.1 and 4.3 are made considerably more straightforward when dealing with a number of smaller fully connected graphs than would be the case with a single network of largely unconnected graphs. Besides which manipulating a single enormous semantic network is considerably more taxing on computing resources than dealing individually with a number of smaller concept graphs.

The partitioning of WordNet can be described by imagining WordNet as a two dimensional mass of nodes and links. The partitioning process could then be thought of as the act of selecting one of these nodes as a root node and lifting it clear of the surface. All nodes connected to this root node by hierarchical links are likewise pulled clear of the surface, leaving us with a hierarchical concept graph, (HCG). Within this newly constructed HCG there are links that traverse the hierarchy, (the Part-of, Member-of and Substance-of links). The large majority of these links connect nodes within the HCG, however, some of them reach down to connect back with nodes on the surface. For now we will ignore these links and imagine we have a single autonomous HCG. The problem then, of course, is to decide what the root concepts for each HCG should be. In the original construction of WordNet, a set of 25 primitive concepts was decided upon following an analysis carried out by Johnson-Laird, [Mill90b]. An initial attempt at constructing HCGs thus entailed using 23 of these 25 primitive concepts as root concepts for individual HCGs. The other two primitive concepts resulted in very small HCGs being generated, and were thus discarded. The root synsets were :

- | | |
|-----------------------------|----------------------------|
| - { Action, Act, Activity } | - { Animal, Fauna } |
| - { Artifact } | - { Attribute, Property } |
| - { Body, Corpus } | - { Cognition, Knowledge } |
| - { Communication } | - { Event, Happening } |
| - { Shape } | - { Person, Human_Being } |
| - { Feeling, Emotion } | - { Food } |
| - { Group, Collection } | - { Location, Place } |
| - { Natural_Object } | - { Natural_Phenomena } |
| - { Possession } | - { Process } |
| - { Plant } | - { Quantity, Amount } |
| - { Relation } | - { State, Condition } |
| - { Substance } | |

An evaluation of these HCGs then took place. The evaluation was based on the coverage of the original WordNet noun hierarchy, and overlaps found between the HCGs. In terms of coverage it was found that 97% of the original WordNet concepts were retained with the above set of HCGs. The overlapping of HCGs was determined by comparing the concepts in each HCG against the concepts in all other HCGs. The basic finding was that the 'Food' HCG was found to completely be contained in the 'Substance' HCG and the 'Communication' HCG was found to be completely contained in the 'Relation' HCG. Detailed results of the analysis can be seen in Appendix A.

A second attempt at arriving at a set of root nodes for HCGs involved a search for concepts that have no IS-A parent node. The following synsets were subsequently chosen as roots for HCGs :

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

Basically many of the root concepts suggested in the previous attempt were collapsed into the { Entity } root concept¹⁹. This set of root nodes was also evaluated in terms of their coverage of nouns in the WordNet database and the degree of overlap between the resulting HCGs. There was 100% coverage of all WordNet concepts and details of the amount of overlap between HCGs can also be found in Appendix A. This arrangement decreases the degree of HCG overlap while increasing the coverage of concepts. The only concern is with regard to the size of the 'Entity' HCG, which is very large. Problems could be envisaged with manipulating a structure of this size. Nevertheless we decided on this arrangement as the one to be used for our WordNet derived KB.

The actual HCG construction process involves several passes through the WordNet noun data file. On the first pass all concepts below the HCG root concept are extracted along with their pertinent information. A second pass is required to resolve all relational pointers. It must be remembered that pointers still point to the WordNet concepts and must be made to point to the corresponding concepts within the HCG. The extraction process is complicated by the existence of non-hierarchical link types. As was mentioned above, it is possible that some of the non-hierarchical links within a HCG may point to WordNet nodes that were not extracted for the HCG. Initially we concentrated on ensuring all nodes of the original network were members of some HCG and we dealt with all of these 'stray' HCG links by simply deleting them. Although information was being lost it was believed that HCGs were relatively independent anyway, and therefore not many of these 'stray' links existed. However, following some testing of the system we were prompted to investigate just how 'few' of these links there were. There turned out to be in the order of ten thousand. Consequently, we recognised that allowance would have to be made for them.

Extending individual HCGs to allow for inter-HCG links would involve adding an extra field to indicate whether a link was internal or external. In addition the HCG construction operation would involve an extra pass to resolve byte offsets of external links. Instead we opted for the alternative of using inter-HCG pointers as an

¹⁹ Described as 'something having concrete existence; living or nonliving'

index to a central table which in turn directed them to the appropriate HCG and the offset within that HCG. This required changes to the HCG builder so that all unresolved non-hierarchical pointers were retained. The offsets pointed to by these pointers were altered so the first digit is changed to a '9', since this digit is never used anyway, (its always '0'), there is no effect to the actual offset. Hierarchical unresolved pointers are already taken care of because the nodes they point to are duplicated in all involved HCGs. The retained pointers effectively gave us all the WordNet nodes that are involved in a cross HCG link. These pointers were then used to construct a lookup table of the form :

WN_offset Num_HCGs { HCG_ID HCG_Offset }

Any relational links beginning with a '9' could then be used as an index into this table to locate the corresponding HCG offset in some other HCG.

The format of the HCG file is as follows :

Byte_offset Con_num Num_wrds { Word_form Sense_num } Num_ptrs { Ptr Ptr_offs } | Gloss

This leaves out some of the original WordNet information not relevant or useful to our system, and introduces the concept number field, (useful for possible HCG expansion).

Having decided upon the set of HCGs, it was necessary to construct indexes to allow quick access to concepts in the knowledge base of HCGs. A two tier indexing mechanism was set up whereby an overall index pointed to a HCG index which in turn pointed to byte offsets within actual HCGs. Given the certain overlap found between HCGs, the super index entry for particular concepts will point to more than one HCG index file. The format of the super index is thus :

Concept_name Num_senses { HCG_IDs }

The format of the HCG index file is as follows :

Concept_name Num_senses Polysemy { HCG_Data_file_offset }.

The concepts are sorted alphabetically on the Concept_name field.

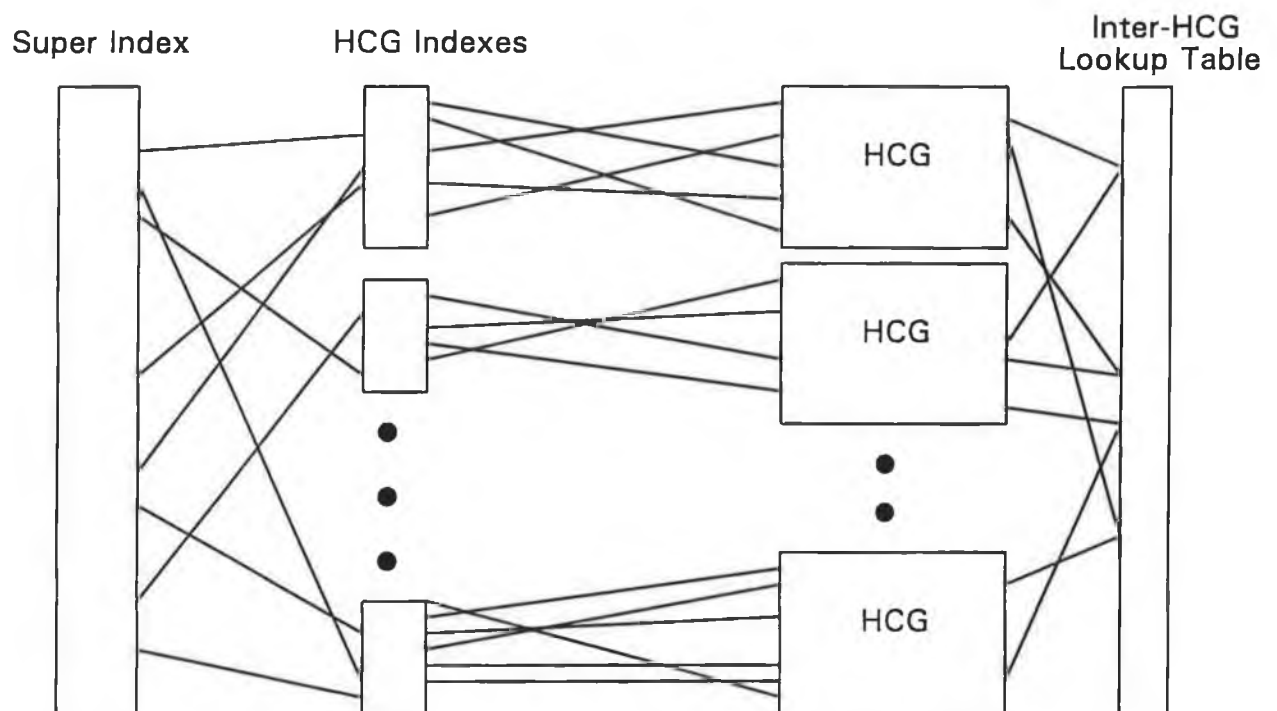


Figure 4-1 Indexing of concepts in KB

4.2 Extension to HCG synsets

Having described the structure of our WordNet derived KB in the previous Section, we can now describe the first of our extensions. It is well known that synsets near the top of a HCG are quite broad and general in meaning and synsets near the leaf nodes of HCGs are quite specific in meaning. There is, however, no field in our KB reflecting the broadness or information content value of a synset. Although it

may not yet be clear how useful a measure such as this will be, it will be shown in later Sections and Chapters how such a value can be used in as diverse a set of activities as word sense disambiguation, automatic link weighting, and concept similarity estimation. The discussion here concentrates on how such a measure is arrived at and how our KB is extended to include it. Section 4.2.1 describes some work carried out by Resnik in this area; Section 4.2.2 describes an initial implementation of this work and discusses some of its failings; and finally, Section 4.2.3 presents details of an improvement on this work.

4.2.1 Calculating a value for Information Content

Resnik in his thesis, [Resn93a], describes a measure for the specificity or information content value of a WordNet noun synset. He viewed each noun synset as a class of words. The class is made up of all words in the synset as well as words in all directly or indirectly subordinate synsets. The information content value of a class is then defined in terms of the probability of occurrence of the class in a large text corpus :

$$Info_content(C_i) = -\log(P(C_i)) \quad (1)$$

where $P(C_i)$ is the class probability of class i .

In order to define the probability of a class it is first necessary to 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 $\{c | w \in words(c)\}$; i.e., it comprises of all the classes in which the word w is contained, regardless of the particular sense of w . From these two definitions the frequency of a class is defined as :

$$Freq(C_i) = \sum_{w \in words(c)} \frac{1}{|classes(w)|} \times Freq(w) \quad (2)$$

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

$$P(c) = \frac{Freq(c)}{N} \quad (3)$$

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

Resnik used 800,000 noun occurrences from the Associated Press Newswire corpus to calculate class probabilities. In his implementation he ignored collocations. The following Section describes an implementation of this information content measure using an on-line extract of the Wall Street Journal, (WSJ), as the text corpus.

4.2.2 Implementation Details

In our implementation, 11 million noun occurrences from an extract of 276 megabytes of the Wall Street Journal (WSJ) corpus were used for calculating class probabilities. The text of the corpus extract was first processed to remove headers and trailers and then fed through the RUCL part of speech tagger, [Karl89], (see Section 7.1.1 for a more detailed discussion of the RUCL tagger). This analysis program takes text files as input and produces a list of each word token with its syntactic information. This syntactic information includes the morpho-syntactic base form of the word (the stemmed word), its allowable grammatical categories and possible syntactic functions within the phrase in which it occurs. However, for the purposes of this research we were just interested in detecting noun occurrences. All words classified as being *Head Nouns* or *Noun Modifiers*, (see Appendix E), were extracted and used to build an

inverted file with an entry for each word and a count for the number of occurrences in the text corpus²⁰. Certain peculiarities of the syntactic parser were noted at this stage, (see Chapter 7 for further discussion). For instance, a lot of word occurrences were found to have full stops and commas as their last characters. Steps were taken to allow for these 'bugs' in the parser.

Having converted the corpus extract to the above format, the following algorithm was followed :

```

For each word, w, in the WSJ index
{
  Locate the equivalent HCG synset(s) i.e. Classes(w)
  For each element, c, of Classes(w)
  {
    Trace to the root - to get the synsets where c is a member of their
    Words() function
    Place the offsets of these synsets in the array Ci_offsets
    For each element of Ci_offsets
      Increase its Frequency value by  $\frac{1}{|Classes(w)| \times Freq(w)}$ 
    }
  }
}

```

As can be seen from the algorithm all senses of words are included when attempting to match WSJ nouns with HCG concepts. If, as is likely, the WSJ word is not found to match any HCG concepts, another attempt is made to find a match by sending the WSJ noun through WordNet's morphological processor, (although the RUCL parser reduces words to their base forms it often leaves the plural form if it believes the word refers to a group).

²⁰ It should be noted that these occurrence values could be exaggerated slightly. The RUCL parser tries to disambiguate all words in terms of their syntactic category. In cases where it is not possible to disambiguate fully a word, it may be attributed more than one syntactic category.

The output of this algorithm is a transaction file containing an entry for each HCG synset that was either directly equivalent to a WSJ noun or was superordinate to one that was found to be directly equivalent. Each entry contains the byte offset of the synset and a value for its frequency value, as calculated above. This transaction file was then run against the existing HCG data file and all HCG synsets with an entry in the transaction file had an information content value attributed to it. This information content value is calculated as follows :

$$Inf_content(syn_set_i) = -Log(\frac{Freq(syn_set_i)}{NUM_WSJ_NOUNS}) \quad (4)$$

where syn_set_i is the i^{th} synset, $Freq(syn_set_i)$ is the frequency value for the i^{th} synset (from the transaction file), and NUM_WSJ_NOUNS is the number of nouns in the corpus extract, (11,000,000). As a result of this process the format of the HCG data file is changed as follows :

Byte-Offset Info_content Con_num Num_words { Synonyms } (Num_ptrs { Ptr_offset }) | (Glossary)

At this stage all that was left to do was resolve all pointer offsets to take account of the new byte offsets due to the introduction of the information content field. This was handled by the multi-pass procedure used to construct the original HCGs.

4.2.2.1 Analysis of results

Appendix B contains detailed results of the processing. Of particular interest is the fact that on average just under half the synsets of the HCGs didn't receive an information content value. Particular characteristics of these synsets include the high percentage of collocations, the comparatively low polysemy²¹ values and the fact they

²¹ The value for the polysemy of a HCG concept is a count (from the Collins Dictionary of the English Language), of the number of senses of that word when it is used as a noun, verb, adjective, or adverb.

tend to appear deep in the concept graph hierarchies. All these results comply with what would be expected. The set of synsets that did receive an information content value can be split between those that contain words that were directly found in the text corpus and those which received a value because they are superordinate to the directly found synsets. The two groups are referred to as the directly found and indirectly found synsets. Of course, these two sets are not mutually exclusive. It is possible for a synset to be amongst the directly found set and also to receive a frequency value indirectly from a directly found subordinate synset, (the algorithm to determine frequencies of classes or synsets would add together these two values to get an overall frequency). In terms of results for these two sets it is apparent that only a small proportion of found concepts receive their information content values by indirect means alone. The proportion of found concepts, direct or indirect, which are collocations is considerably lower than amongst the unfound concepts.

The most important results of the implementation just described are that a large proportion of synsets did not receive an information content value and that a high percentage of these synsets contain collocations. The fact that Resnik ignored collocations in his computing of information content values now appears to have been ill-advised. The extent of this neglect can be seen more clearly in the fact that collocations make up approximately 45% of all WordNet concepts. In an attempt to improve this situation it was decided to attempt another implementation in which there would be provision for collocations. This second implementation is described in the following section.

4.2.3 Improving Results

The previous Section described an implementation of Resnik's approach to calculating class probabilities, and as a result of this implementation certain shortcomings of Resnik's approach were highlighted. Foremost among these caveats was the fact that well over half the synsets in WordNet didn't receive an information

content value²². Given the fact that the information content value of a synset is defined in terms of the information content values of its subordinate synsets, the absence of an information content value for a given synset affects both itself and its superordinate synsets. The most obvious explanation of this regrettable situation would seem to be the omission of collocations (which account for approx. 45% of WordNet concepts) in the process of calculating information content values. As such, this Section describes how we handled collocations and analyses the results obtained.

Section 4.2.3.1 describes our treatment of collocations. Section 4.2.2.1 gives an analysis of the results of our approach and conclusions are presented in Section 4.2.4.

4.2.3.1 Information content values for collocations

Collocations are simply multi-word phrases, in other words bigrams and trigrams, and if the text was parsed to produce an inverted file in terms of bigrams and trigrams, as well as individual words, it would be possible to include collocations in the calculation of class probabilities. The RUCL syntactic tagger does actually parse the text for collocations as well as individual words, a lexicon of collocations is looked-up during the parse. Unfortunately, however, this lexicon is quite small, containing only a small fraction of the collocations in WordNet. One option for dealing with collocations was then to extend this lexicon with all the collocations in WordNet and to re-parse the WSJ text. The only problem with this approach was the speed of the RUCL parser, (on our computer system it took four hours to process one megabyte of text). Another month would be required to re-parse the text. The only other alternative was to use the existing parsed text and to extract the collocations from this text. Fortunately, the parse preserved the sequence of the text and it was possible to 'plough' through the text looking for collocations.

²² It found that on average over 50% of the synsets in any given HCG did not receive an information content value. This was using a text corpus extract of 11 million noun occurrences. We suspect Resnik recorded a far higher percentage of synsets without a value as his corpus extract contained only 800 000 noun occurrences

```

FOR all WSJ Files
  WHILE not eof(WSJ_file)
    read a word from the WSJ file
    IF wsj_word is a noun
      FOR all HCGs
        IF wsj_word is first word of any collocation in that HCG
          Read in preliminary list of collocations that have
            wsj_word as first component
          REPEAT
            Read in subsequent wsj_word
            Check subsequent elements of preliminary list
          UNTIL there is no longer a match in the preliminary list
          IF any collocation completely matched
            Take largest matched collocation
            Write it to matched collocation file for this HCG
          ENDIF
        ENDIF
      ENDFOR
      IF wsj_word not found in any HCG
        write wsj_word to new WSJ file
      ENDIF
    ENDIF
  ENDWHILE
ENDFOR
FOR all HCGs
  create inverted file for located collocations
ENDFOR

```

The procedure was complicated somewhat by the fact that collocations can have a variable number of component words, (found to vary between 2 and 6, see Table 4.1), and although a collocation had to begin with a noun it could contain words of any syntactic category (e.g. 'chamber_of_commerce', 'department_of_transport', etc.). The underlying premise of the collocation locator was to locate the largest collocation possible. As such, if the co-occurring pair of 'chief' and 'executive' were found in the text, thus allowing for the KB collocation 'chief_executive' to replace them, the next word in the text will still be examined to determine if the KB collocation 'chief_executive_officer' could be applied instead.

The output of the process was a 'collocation file' for each HCG and a new set of WSJ files. The collocation files were combined with intermediate files in the original process of determining information content values, and in so doing, synsets containing collocations obtained an information content value. It was necessary to create new WSJ files since component words of collocations found in the WSJ corpus extract could not again be used in the calculation of frequency values for individual nouns.

4.2.3.2 Analysis of Results

This Section presents the results of our treatment of collocations. Before presenting results, however, there will be an overview of the extent of collocation occurrences in each HCG. Table 4.1 presents details of the number of concepts, the number of collocations, and the break-up of these collocations in terms of the number of component words, for each HCG. Figure 4-2 presents a broader, but more readable view of the percentage of collocations in individual HCGs.

HCG	Total Cons.	Total Collocs.	2-word	3-word	4-word	5-word	6-word
Abstraction	11781	2933	2548	339	41	4	1
Act	7623	1648	1463	156	27	2	0
Entity	43950	22148	20155	1843	123	24	3
Event	1311	322	300	21	1	0	0
Group	11318	6104	5900	185	13	3	3
Location	2664	1032	697	265	58	12	0
Phenomenon	1190	475	430	38	6	1	0
Possession	964	434	385	43	6	0	0
Psych._feature	3837	902	779	102	17	4	0
Shape	688	178	161	14	1	2	0
State	2506	644	570	61	12	1	0
Totals	87832	36820	33388	3067	305	53	7

Table 4-1 Distribution of collocations in HCGs

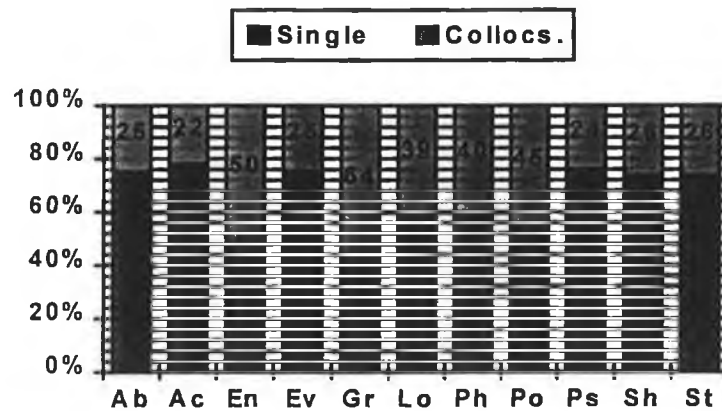


Figure 4-2 Single Vs Collocation concepts

Of particular interest from the above chart is the fact that the HCGs, *Group* and *Entity*, have a particularly high percentage of collocations. If we now view the original results of implementing Resnik's similarity measure, i.e. with no treatment of collocations :

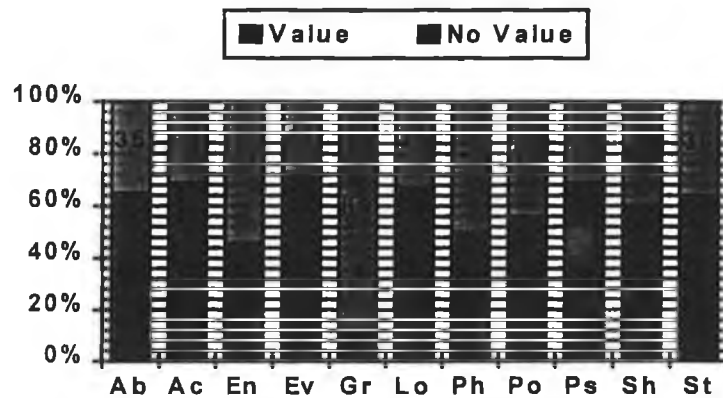


Figure 4-3 Synsets with Information content Values Vs. those without

it can be seen that the *Entity* and *Group* HCGs contain the highest percentage of synsets without an information content value. These results provided our motivation for attempting a treatment of collocations in the process of obtaining information content values for HCG synsets.

Section 4.2.3.1 has already described the process employed to include collocations in information content value calculations. The results of this process are presented below. Figure 4-4 presents the number of unique collocations found for each HCG as a percentage of the total collocations for that HCG.

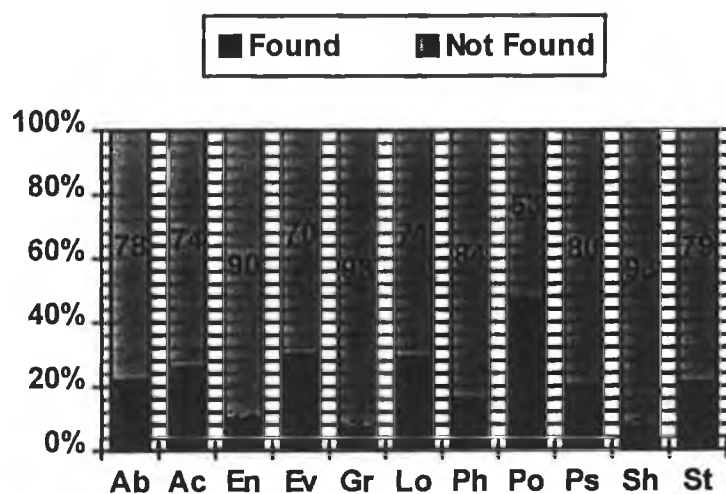


Figure 4-4 Percentage of collocations found in WSJ extract

Upon examination of these results we see that overall the percentage of collocations found to occur in the WSJ extract is not very high, relative to the total number of collocations in WordNet. Reasons for this could include :

- The domain of the WSJ is not sufficiently broad to include more than a small percentage of WordNet collocations
- The WSJ extract used was too small
- WordNet collocations are often very domain specific.

The first point is true insofar as the subject of much of the text in the WSJ has a distinct business/financial emphasis. However, the WSJ is still one of the most general large text corpora to be found on-line. The third point would also seem to be responsible for the comparatively small percentage of collocations found. Looking at some of the collocations :

'genus_aeonium'	(Group)	'genus_agam'	(Group)
'genus_aeyyplopithecus'	(Group)	'genus_aeygupius'	(Group)

it can clearly be seen that they are technical terms and certainly very domain specific. As a result it is very unlikely they would be found in even a general domain text corpus. In light of the above discussion the second possible reason for the result obtained appears unlikely. However, in order to verify this the procedure to locate collocations was rerun with different sized WSJ extracts.

HCG	Total Collocs.²³	1 WSJ.	20 WSJs	40 WSJs	276 WSJs
Abstraction	2879	99	376	492	616
Act	1624	45	246	325	422
Entity	21367	184	1189	1604	2093
Event	321	8	53	71	95
Group	6071	95	277	330	378
Location	1012	52	200	248	291
Phenomenon	471	6	43	54	73
Possession	434	53	147	168	203
Psych._feature	892	19	105	145	180
Shape	178	0	5	9	9
State	634	12	81	105	133
Totals	35883	573	2722	3551	4493

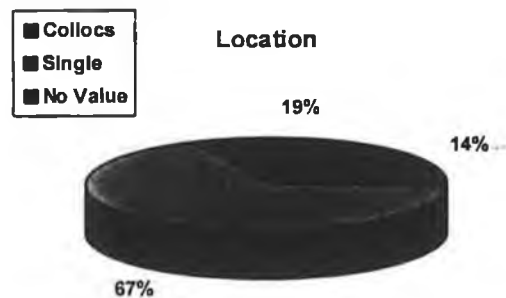
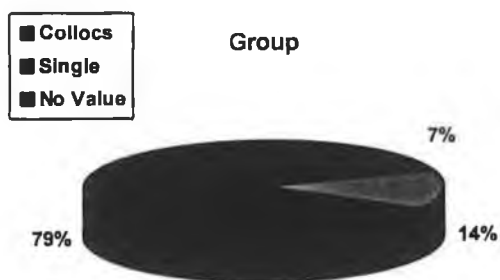
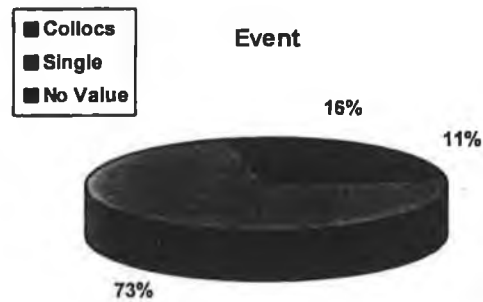
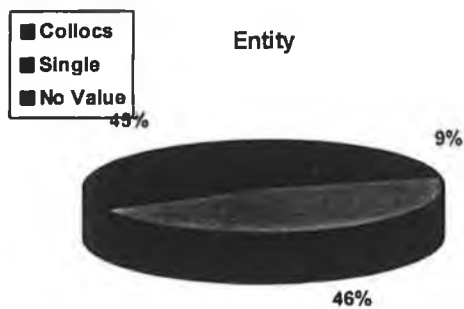
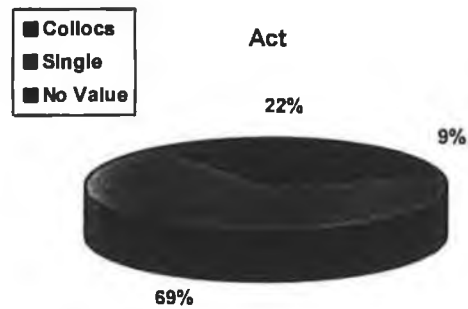
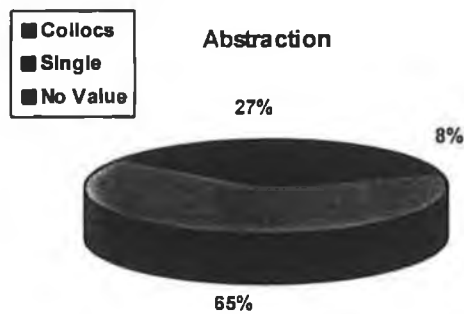
Table 4-2 Unique collocations detected in different sized WSJ extracts

The WSJ corpus available to us was made up of 276 files, each about one megabyte in size. Three re-runs of the collocation detecting process were set up, one with just one WSJ file, one with 20, and the third with 40. Table 4.2 compares the results of these three re-runs against the results of the original run with 276 WSJ files for all HCGs. As can be seen from these results the majority of detected collocations were found in the first 40 WSJ files, (just 15% of the complete text). The remaining 85% of the text only served to increase the number of occurrences of the already located collocations. It is clear that very little benefit would be derived from using a larger excerpt of the WSJ corpus.

The most important result of the procedure to detect collocations is presented below. The 11 pie charts of Figure 4-5, illustrate the increase in the number of synsets that received an information content value as a result of the treatment of collocations.

²³ This is different from the value in Table 4.1 because it is the total unique collocations, i.e. ignoring polysemous meanings

The segment labelled *Single* represents the percentage of synsets that received an information content value when collocations were ignored, the segment labelled *Collocs* represents the set of collocations that received an information content value as a result of including collocations, and the segment labelled *No Value* represents those synsets that still have no information content value.



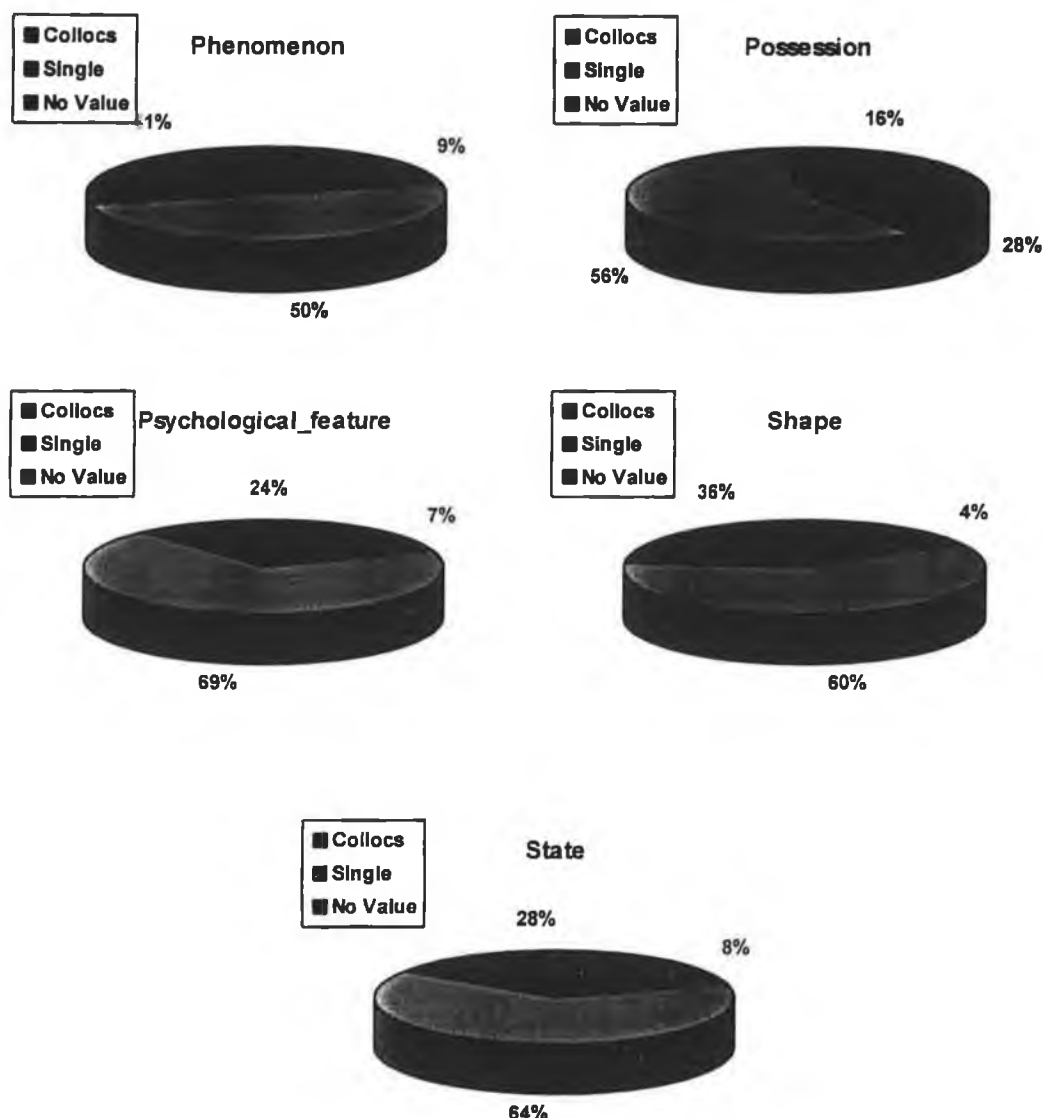


Figure 4-5 Inclusion of Collocations in Information Content Calculation

As can be seen from the charts there is an appreciable increase in the number of synsets receiving information content values as a result of the treatment of collocations. The best result can be seen for the *Possession* HCG. Particular characteristics for this HCG include the fact that 44% of its concepts are collocations. However more importantly, 47% of these collocations were found in the WSJ extract. The worst result, in terms of the number of synsets without an information content value, is the *Group* HCG. Again the particular characteristics of this HCG are as

would be expected; 64% of the concepts are collocations and only 7% of these were located in the WSJ corpus extract.

There is no direct linear relationship between the number of collocations found and the increase in synsets receiving information content values. This is both due to the nature of information content value calculation and to the fact synsets and not individual concepts receive information content values. Dealing first with the nature of the information content value calculation we see that synsets can receive their information content values directly or indirectly. If a concept within a particular synset is found to occur in the text corpus then that synset gets an information content value directly. However, synsets also receive information content values because they are superordinate to some directly found synsets such synsets are referred to as the indirectly found synsets. It is possible that some of the synsets that directly received their information content values as a result of the collocation detection process could provide information content values to superordinate synsets which otherwise would not have received a value. Unfortunately, this situation is rare because the higher familiarity of superordinate concepts makes it more likely that they will have acquired an information content value directly anyway. However, it should be noted that the information content value of any synset is made more accurate by the inclusion of as many subordinate synset information content values as is possible.

The second reason why there is not a linear relationship between collocation detection and an increase in the number of synsets with an information content value has, quite simply, to do with the fact that synsets and not individual concepts receive information content values. If we have the following synset :

{ *savings nest_egg* }

suppose the collocation *nest_egg* is discovered by the collocation detection scheme, this in turn guarantees this synset will receive an information content value. However, we might not have an increase in the number of synsets found because another member of the synset, (*savings*), may already have been discovered in the text by the

single term detector. However, as before, the information content value is more accurate as a result of the contribution due to *nest_egg*.

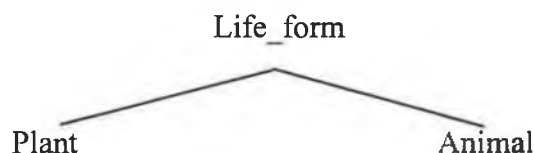
4.2.4 Conclusions on the Information Content value Extension

In the previous Sections there was a description of a procedure to extend our KB to include a field which measured the information content value of a synset. Initial research reimplemented work carried out by Resnik in [Resn93a]. However, it was found that some of our synsets did not receive an information content value using this approach. Section 4.2.4 presented the results of a procedure to improve upon Resnik's approach to calculating the information content of classes of words. The improvement was as a result of a process to include collocations in the calculation of information content values. An improvement upon Resnik's original implementation had already taken place due to the use of 11 million noun occurrences as opposed to the 800,000 he used. However, the results here show a further increase in the number of synsets receiving an information content value. As was stated a number of times throughout this Section the improvement due to the inclusion of collocations can be seen in both the increased accuracy of existing information content values as well as in the expansion to the set of synsets with an information content value.

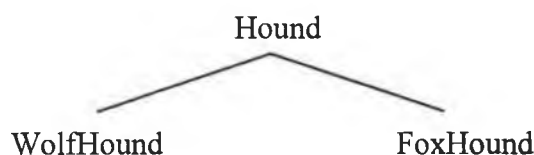
It would be interesting to re-run the collocation detection process on a more general text corpus than the news-specific WSJ corpus. From the results here it is apparent that a large text corpus is not really required and that a corpus with a more general domain would probably produce even better results. However, it is also quite obvious that it is highly unlikely that a situation would arise where occurrences of all collocations, and for that matter single term concepts, could be found.

4.3 Weighting HCG Relational Links

A second, and perhaps more obvious extension to our WordNet derived KB, is the weighting of the semantic links within HCGs. When one imagines the concept space of a HCG it is reasonable to assume all concepts are not equidistant from each other. Take for example the concepts:



Near the top of a HCG as compared with something deep in the hierarchy, as in:



It is clear that the concepts deeper in the HCG are conceptually closer than those near the top. This fact should be reflected in the weighting of links. However, unlike the concept graphs of other researchers, ([Gins93], [Rada89], [Kim90], and [Lee93]), those created for our research 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].

In Section 4.3.1 the work carried out by Sussna is reviewed. Section 4.3.2 presents our proposed improvements to Sussna's approach. Finally Section 4.3.3 briefly discusses how we implemented our automatic weighting mechanism and presents our conclusions on link weighting.

4.3.1 Previous Weighting Strategies

Sussna, in [Suss93], hypothesised that the value for the weight of a link in a concept graph is affected by the following :

- (a) the depth in the HCG - conceptual distance shrinks as one descends a hierarchy
- (b) the density of the HCG at that point - conceptual distance is less in a dense part of the network

The first point can be seen from the 'Life_form' and 'Hound' example of the previous Section, however, the second point is not so clear. In a personal communication with Ellen Voorhees, she commented :

'... need to be careful about how you define link weights. Different parts of WordNet are denser than others (e.g., the plant hierarchy), and one IS-A link in a dense part of the hierarchy represents a much smaller conceptual distance than one IS-A link elsewhere.'

The plant Section of WordNet is a very dense. Individual nodes having up to three or 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.

Sussna measures the density of the network around the link being weighted by counting the number of links of that type between the source and destination nodes. He restricted this measure to fall between the scales $\max_r$ and $\min_r$ as follows :

$$W(X \xrightarrow{r} Y) = \max_r - \frac{(\max_r - \min_r)}{nr(X)} \quad (1)$$

where X and Y are two adjacent nodes, $\xrightarrow{r}$ is a relation of type r , $\max_r$ and $\min_r$ are the maximum and minimum weights possible for a relation of type r , and n_r is the number of relations of type r leaving node X . Sussna addressed the depth input by simply dividing the value returned by the above local density calculation by the depth of the link in the network. Finally, given the fact that each edge in WordNet is really a pair of inverse relation links :

- Holonym / Hypernym (IS-A / Has Kind of)
- Meronym Part-of / Holonym (Part-of / Has Part)
- Meronym Member-of / Holonym (Member-of / Has Member)
- Meronym Substance-of / Holonym (Substance-of / Has Substance)

by averaging the weights from a link pair Sussna arrived at the edge's undirected weight. The overall weight of a link can thus be expressed using Sussna's formalism as follows :

$$W(X, Y) = \frac{W(X \xrightarrow{r} Y) + W(Y \xrightarrow{r'} X)}{2d} \quad (2)$$

where $w(X \xrightarrow{r} Y)$ is as above, r' is the inverse relation type of r , and d is the depth of the deeper of the two nodes X and Y .

In [Suss93] the results of a set of experiments on the use of WordNet in word sense disambiguation are detailed. In the experiments WordNet is used as an enormous semantic network. When an ambiguous word is encountered an attempt is made to disambiguate it by choosing the sense, in WordNet, which is conceptually closest to its context, which is captured by the words surrounding it. The results of the experiments were tested against the performance of humans in the same situations.

In the course of the experiments, the methods of weighting WordNet relational links were varied and the results compared. In the first variation the depth scaling was

removed, the ensuing results were quite bad, indicating that relative depth scaling is an important factor in weighting links. The second variation involved removing the local density estimator, but surprisingly there was no significant change in the results. Motivated by these findings we worked on developing an improved automatic weighting mechanism.

4.3.2 New Link Weighting Mechanism

Details of our proposed improved automatic link weighting mechanism are presented in this Section. As with Sussna's approach, we base our weighting mechanism on the observations that links in a dense part of the hierarchy should account for a smaller conceptual distance than those in a less dense region, and that conceptual distance shrinks as one descends a hierarchy. A third factor we introduce is :

- the strength of connotation between parent and child nodes.

The point can be illustrated by the following diagram, (where the numbers represent the node's information content values):

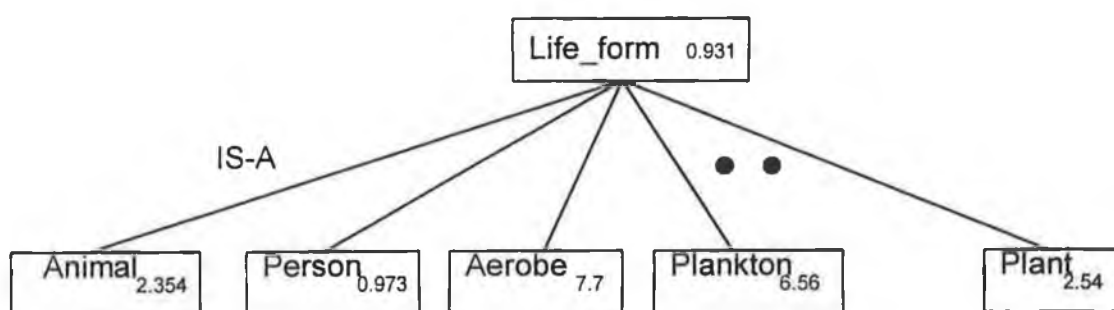


Figure 4-6 KB Extract

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*. The

strength of connotation of the link being weighted can be estimated as a function of the information content values of the source and destination synsets and of their sibling synsets. Strong parallels can be drawn here between Tversky's definition of the salience of a feature and the relationship between the parent and children nodes, [Tvers77] (refer to Section 5.1.3 for further discussion). A possible formalism for the strength of connotation of a link is represented as follows :

$$St_i = 1 - \frac{\sum_n \frac{1}{info_cnt}}{\frac{1}{info_cnt_i}} \quad (3)$$

where St_i is the strength of connotation of link i , $info_cnt_i$ is the information content of the destination synset and n is the total number of links of this type emanating from the source synset.

As well as introducing the strength of connotation factor into the link weighting mechanism, we also proposed changes to the local density and depth scaling factors. One of the main findings of Sussna's experiments with his weighting mechanism was the fact that his method of estimating local density was not good. Consequently, we investigated other methods of determining the density of a concept graph at a particular point. Under our proposal the local density of a link connecting a source node to a destination node is estimated by :

$$Den(X) = \frac{Sour_fan + (par_fan + sib_fan + des_fan) \frac{d}{d+1}}{Num_fans} \quad (4)$$

where d is the depth of the deeper of the source and destination nodes, $sour_fan$, par_fan , sib_fan and des_fan are the fanouts of link type r for the source, parent(s) of the source, sibling(s) of the source and the destination node respectively, and Num_fans is the number of the above fanouts present for a given source-destination pair. Each fanout is weighted according to its perceived importance. At present, the

fanout between the source and destination nodes is given a weight of 1.0 and all other fanouts are given a weight equivalent to $d/d+1$. This new density estimator replaces $nr(X)$ in (1). $Den(x)$ is intuitively more sensitive to the density of a link type than $nr(X)$. Using the sample HCG in Figure 4-7, under Sussna's weighting algorithm the weights for the three links at point A would be the same as those at point B.

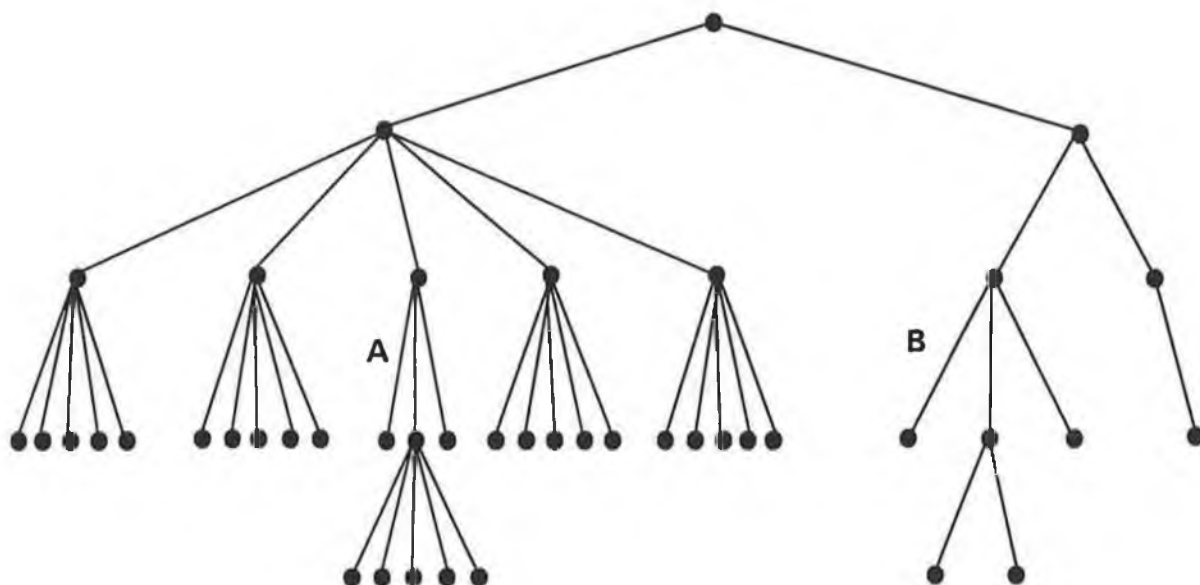


Figure 4-7 Local Densities in HCGs

However, the hierarchy at point A is obviously much denser than at point B, the weighting scheme just described takes this fact into account and assigns a different set of weights to the links at points 'A' and 'B'.

Having changed the local density estimator we then began thinking of possible improvements to the depth scaling factor. One weakness of Sussna's depth scaling factor is apparent in terms of its dependence on the structure of a HCG. In particular, it relies on the assumption that concepts that would be envisaged as being at the same level of abstraction are at equal depths from the root. However, this is not always the case. Using the example of 'horse' and 'cow', (both being regarded as being of the same level of abstraction), the node for 'horse' is 10 levels from the root, taking 'entity' as the root concept, and one for 'cow' is 13 levels deep. As such, because there is a

large body of information in WordNet for one concept relative to another, the weightings for the link from that concept are unfairly penalised.

One possibility of improving this situation may be to use the information content values of nodes, (as described in Section 4.2), in the depth scaling process. It would seem reasonable to assume that concepts at the same level of abstraction would have comparable information content values. Such a depth scaling factor would not be as much at the mercy of structural disparities of our WordNet derived HCGs as one just using the depth from the root. However, it was not at all obvious how we could introduce information content values into the depth scaling. Direct use of the information content values, as with the depth in a HCG, would not be expected to give good results, given the small ranges within which they operate. It was thus decided to set up a fixed width scale within which information content values could fall. Synsets falling within the same level of the scale could be said to be at the same level of abstraction. The question, of course, is how to establish these levels. Following some experimentation with information content values it was decided to use the standard error of information content values from the mean as the level width. The depth scaling factor could be represented as :

$$Ds = ABS(\frac{info_cnt}{\delta} + 1) \quad (7)$$

where *info_cnt* is the information content value of the deeper of the source and destination nodes of the link being weighted and δ is the standard error from the mean for the information content values of the HCG in question.

4.3.3 Conclusions on Weighting

In the previous Sections there was a description of a procedure to automatically weight the relational links our KB. Based on work carried out by Sussna, and to a lesser degree Botafogo, we weighted a link depending on its depth in

the HCG, the local density of surrounding links and the strength of relation between the source and destination nodes. The actual implementation involved a triple pass through the HCG and the use of a variation of the byte offset resolution process used in the HCG construction process. The format of the HCG files as a result of the weighting process is :

Byte_offset Con_num Num_wrds { Word_form Sense_num } Num_ptrs { Ptr link_weight Ptr_offs } | Gloss

Seven configurations of the KB were created as a result of different combinations of the weighting strategies detailed in Section 4.3.2. Each of these configurations were evaluated in a set of experiments described in Section 5.3 of the next Chapter.

4.4 Summary

This Chapter contained a description of how we constructed our KB. The main deriving factor was the body of information contained in WordNet. The organising structure of this information was changed from a massive semantic network into a number of component hierarchical concept graphs. With this organisation, speed of access can be improved and many of the problems associated with handling large information structures can be avoided.

As well as the information contained in WordNet, additional fields were added in the construction of our KB. The first of these extensions was a field which gave individual synsets in our KB a measure of their relative information content value. A large text corpus was used in the calculation of this field. The extension involved the weighting of the relational links between nodes. Links were weighted relative to how closely related the synsets they connected were.

In the following Chapter there is an account of the second step in the development of our knowledge based semantic information processing system. This step is concerned with the development of an estimator of the semantic similarity between words. An overall evaluation of both components of our system will be presented in Chapter 9.

Chapter 5 - Similarity

5. Introduction

Up to now we have concentrated on developing WordNet into a knowledge base to be used in an information processing application. In an information retrieval task we could use this knowledge base as a controlled vocabulary for the representation of both the information request and the information being searched. This effectively eliminates problems caused by homonyms, words with the same form but a different meaning. If the information request includes the word 'bank' we will know whether it is the commercial meaning of bank or the river bank meaning. A complete retrieval system using pattern matching between word variants as the comparison process could now be built on top of this KB. However, problems would still be posed by relevant information being described by words not used in the information request but rather by related words. Unfortunately, it would appear that problems of this nature are far more likely than those caused by homonyms, [Bate86]. It would thus seem necessary to attain a more complete treatment of the semantic issues in information processing than is already the case. Given the richness of natural language and in particular, the multitude of ways in which the same thing can be described, pattern matching as a comparison process is simply too black and white. It is thus proposed to replace pattern matching with a conceptual similarity estimator which will use the semantic information in our KB to rate the similarity between words. This will increase the power of our information processing system so that will recognise related terms such as truck and lorry, watch and clock, car and automobile, etc..

The details of the semantic similarity estimators used in our system are described in this Chapter. Section 5.1 briefly summarises previous related work in the area of similarity estimation. Section 5.2 describes the two similarity estimators

employed in this research. In Section 5.3 there is a description and evaluation of the results of an evaluation of these and other similarity estimators, using human judgement as a baseline. This Chapter concludes with a summary of developments so far and an outline of the applications of our system that are presented in Chapter 6.

5.1 Related Research

Similarity is one of those philosophical issues that has been deliberated over by a number of different disciplines for centuries. Most of these studies have concentrated on trying to understand the cognitive process behind how humans rate the similarity of objects²⁴. However, it is important to note our interest is not in theories of how humans perform similarity comparisons but simply in arriving at a method of simulating human judgements of similarity²⁵. In this section there is a review of three existing approaches to modelling conceptual similarity. Two of the approaches are based on two contemporary models of how human memory is organised namely, the set-theoretic model, [Meye70] and the geometric model, [Quin69]. According to the set theoretic model ‘.. concepts such as robin, bird, and animal are represented by a set of elements where these elements may be exemplars, attributes, subsets or supersets of the concepts.’, [Rips73]. Similarity in this model is measured by the degree of contrast and overlap between the sets of elements representing concepts. According to the network model, human memory is organised in terms of a network of concepts connected by labelled relations and the retrieval of information is through a process of spreading activation along the ‘associate’ links. Similarity of concepts is thus measured by the number of links between concepts. A third broad area of research into models of similarity is primarily concerned with the

²⁴ There is no distinction made between rating the similarity of words, of sounds, of pictures, etc.. It is assumed the same cognitive process is involved in all such comparisons.

²⁵ In this goal we assume that human judgements of similarity between concepts can be scaled up to give human judgements of relevance between composite objects such as documents.

synonymy of words. Various researchers have investigated how humans determine synonymy and the degree of synonymy of words.

In the following subsections there will be a more detailed description of each of the models of similarity, as introduced above. The discussions focus on what aspects of each model may be useful in our search for a similarity estimator to replace direct pattern matching. We do not claim that our knowledge base is an accurate model of human memory,²⁶ or even that human memory organisation is suitable for the retrieval task we have set ourselves, [John88]. Our aim is simply to acquire an estimator of word similarity using the resources at our disposal²⁷. The set of synsets, relational links and information content values in our KB. It is hoped by doing this we will overcome the inadequacies of pattern matching between words as a comparison process.

5.1.1 Models of Synonymy

[Rub65], reports on a set of experiments that try to establish the usefulness of the contextual representation of words as an estimator of semantic similarity. In this study contextual representation was estimated by term co-occurrence and human judgement was used as a baseline. A sample experiment to deduce the semantic similarity between two words, A and B, was organised as follows :

- (i) List all words that occur in a set of contexts of A
- (ii) List all words that occur in a comparable set of contexts for B
- (iii) Calculate some coefficient representing the proportion of words in common to the two lists

²⁶ [Rips73] discusses the notion of semantic distance in terms of the network and set-theoretic models of human semantic memory.

²⁷ The only 'resource' available is the knowledge base which is made up of the set of synsets and their information content values as well as the set of weighted relational links that connect synsets together.

They concluded that :

'it may be safely inferred that a pair of words were highly synonymous if their contexts show a relatively great amount of overlap. Inferences of degree of synonymy from lesser amounts of overlap ... are uncertain since words of low to medium synonymy differ relatively little in overlap', [Rub65].

The poor performance of Rubenstein's approach to semantic similarity is unfortunate given it is quite easy to see how the procedure can be carried out by a computer. The only difficulty with automating the approach is the presence of word sense ambiguity in the contexts of words. Rubenstein manually performed the sense disambiguation of contexts but, as will be seen in Chapter 7, automatic sense disambiguation is a very difficult task to automate. Ironically, a direction currently being taken by researchers in the field of word sense disambiguation, [Mill94], is to use manually tagged corpus extracts to obtain contexts of words for use in the determination of the intended senses of ambiguous words. Parallels can also be drawn between Rubenstein's work and the work carried out earlier by Lesk in the area of automatic sense disambiguation, [Lesk86]. So it would seem contextual representations of words as used by Rubenstein for measuring the synonymy of words may find a better application in the process of automatic sense disambiguation.

In a subsequent study Miller replicated Rubenstein's experiments except they estimated contextual similarity using substitutability, [Mill91]. The format of their experiments was as follows :

- (i) Collect a set of sentences using item A
- (ii) Collect a set of sentences using item B
- (iii) Delete A and B from the sentences
- (iv) Challenge subjects to figure out which is which

Unlike term co-occurrence they found that the method of substitutability operated well at both high and low levels of contextual similarity. They concluded :

' .. contextual similarity is related to semantic similarity and is best estimated by tests of substitutability.', [Mill91].

However, unlike the method of term co-occurrence, it is not apparent how the method of substitutability could be used by a machine to estimate the similarity of meaning of two concepts. Certainly there is no obvious way of applying the method of substitutability to estimating semantic similarity using our knowledge base.

Of general interest from Miller's experiments was the fact that they used a subset of the same test data as Rubenstein, a collection of 30 noun pairs, (see Table 5.1 of Section 5.3.1), and they found that ' People are not only able to agree reasonably well about semantic distances between concepts, but their average estimates remain remarkably stable over more than 25 years'. This provides evidence for the suitability of using human judgement as a baseline in the evaluation of computational models of semantic similarity, as we will report in Section 5.3.

5.1.2 Geometric Models

Geometric models of similarity are based on the network model of human memory organisation as introduced in Section 5.1. It is assumed concepts exist as independent nodes in a large semantic network and are linked together by labelled relations. Relation types provide many different dimensions in the semantic network or conceptual space. Shepard, [Shep62a and Shep62b], developed a technique known as multidimensional scaling to identify conceptual dimensions for particular sets of concepts. Multidimensional scaling requires humans to rate the similarity of all pairs of concepts under consideration. This proximity data is then compiled by the multidimensional scaling for all possible concept pairs and is represented in a

3 senses of transport

Sense 1 [10.00]

{ conveyance, carrier, transport } -- (something that serves as a means of transportation)
=> { instrumentality } -- (an artifact that is instrumental in accomplishing some end)

Sense 2 [6.00]

{ transportation, shipping, transport } -- (the commercial enterprise of transporting goods and materials)
=> { commercial enterprise, business enterprise, business } -- (purchase and sale of goods and services)

Sense 3 [2.00]

{ ecstasy, exaltation, transport, rapture }
=> { happiness, gladness, felicity }

4 senses of car

Sense 1 [8.00]

car, auto, automobile, machine, motorcar, motor car
=> motor vehicle, automotive vehicle

Sense 2 [8.00]

car, railway car, railroad car -- (adapted to the rails of railroad)
=> wheeled vehicle -- (moves on wheels)

Sense 3 [1.50]

car, elevator car -- (where passengers ride up and down)
=> compartment -- (a partitioned section or separate room within a larger enclosed area)

Sense 4 [1.50]

{ car, gondola } -- (carries personnel and cargo and power plant)
=> { compartment } -- (a partitioned section or separate room within a larger enclosed area)

2 senses of sedan

Sense 1

sedan [10.00]
=> { car, auto, automobile, machine, motorcar, motor car }

Sense 2 [6.00]

sedan, sedan chair -- (a closed litter for one passenger)
=> litter -- (a chair or bed carried on two poles by bearers)

2 senses of truck

Sense 1

{ truck } [10.00]
=> { motor vehicle, automotive vehicle }

Sense 2 [6.00]

{ hand truck, truck } -- (a frame with two low wheels and a ledge at the bottom and handles at the top; used to move crates or other heavy objects)
=> { handcart, pushcart, cart, go-cart } -- (pushed by a person; may have one or two or four wheels)

2 senses of cart

Sense 1 [10.00]

{ handcart, pushcart, cart, go-cart } -- (pushed by a person; may have one or two or four wheels)
=> { wheeled vehicle } -- (moves on wheels)

Sense 2 [6.00]

{ cart, two-wheeler } -- (has two wheels)

Upon examining these results we see that the correct choices were made with the word 'transport'. Since this word was used as an example of the usefulness of the information content disambiguator in the previous Section there is no further need for discussion. It might be thought that an unusual decision was made in the tying of scores for the railway and automobile senses of car, however, given the overall subject of the article, (perhaps describable as haulage or transportation), the decision might not seem that strange. Interestingly the locality disambiguator chose the railway sense of car and the information content disambiguator chose the automobile sense. When this is further explored it is easy to see why both senses are quite close together in the KB yet the context terms 'cart', and 'bicycle' are considerably closer to the railway sense than to the automobile sense. The final two ambiguous words, truck and cart, are also correctly disambiguated.

7.3 Possible Improvements to Sense Disambiguator

The semantic tagger developed for this research is novel in its approach. Of particular importance is the fact that all ambiguity is retained thus allowing extra scope in the query matching process. Obviously a more rigorous evaluation would be more satisfactory but given it is only one step in the overall IR evaluation of the research system, the time required to perform this evaluation would be difficult to justify. Moreover, the evaluation of the sense disambiguators of other researchers has also been informal. The obvious work load involved in such a procedure has thus far been too great. Although, Sanderson in a recent article, [Sand94], suggests the use of 'pseudo-words' to evaluate sense disambiguators, the use of a WordNet derived KB in our disambiguator rules out such an evaluation. A more feasible and fitting approach would be to use the manually semantically tagged text of the Brown corpus as a test bed. However, we are reasonably confident that the semantic tagger is quite accurate

following a number of tests of the type reported in the previous Section as well as general exposure to its operation in the large scale.

Despite the apparent success of the sense disambiguator, there is still room for improvement. Some of the improvements that appear most promising are discussed below. They can be listed as follows :

- Variable context window size.
- Frozen window approach to the locality disambiguator.
- Use of corpus frequency values to decide dominant terms for the information content disambiguator.
- Use of the conceptual distance similarity estimator as well as, or instead of the information based estimator in the locality disambiguator.
- Improve the method of capturing context by employing a static hot list of the most content bearing words in an article as well as the dynamic window into the text.

All of these variations on the existing semantic tagger would more than likely bring an improvement to results, however, it is hard to say what effect they would have on the running time. A variable context window would provide more of a context for the synonym, glossary and locality disambiguators. The window could vary in size depending on the ambiguity of the word currently being operated on. The degree of ambiguity could itself be estimated by the number of senses of the ambiguous word in question. The obvious down side of this extension is the fact that a variable window operating under these rules would have a definite degrading effect on the running time. The second proposed improvement stems from the work carried out by Sussna. In some of his experiments he froze the senses of ambiguous words in the context window that had already been disambiguated. In other words, given the word being operated on is the middle one, all ambiguous words to the left of this word have their sense frozen. Whether or not this would bring about an improvement in results is, however, questionable. Sussna's experiments, although using a variation of our locality disambiguator as the semantic tagger, actually found a disimprovement with

the frozen window approach. There was, however, a decrease in runtime. The third improvement may not be too difficult to implement. The use of WordNet familiarity values to decide on the dominant sense is a quick and dirty approach. These familiarity values are calculated from the number of different meanings of the word in the Oxford English dictionary. It would be preferable to use the actual count of occurrences of a word in the calculation of information content values, to decide on dominance of a term. Unfortunately, these frequency scores are no longer available and the effort involved in recreating them at this stage would be considerable. Having said this, however, it is again not directly obvious that, having gone to the effort of obtaining frequency scores, a great improvement would result. For the present, it would seem the use of WordNet familiarity scores is a good enough approach. The fourth extension to the semantic tagger speaks for itself.

The locality disambiguator could possibly be improved by using the conceptual distance similarity estimator instead of, or perhaps in combination with, the information based estimator. Up until recently this would not have been possible due to the comparatively slow running time of the conceptual distance estimator. However, a new faster version of this estimator has been developed for the retrieval engine and this could possibly be used by the semantic tagger.

The final extension to the semantic tagger is with regard to improving the capturing of context. At present, context is captured via the moving window. The proposal here is to supplement the context window with a hot list of terms best describing the text. The problem, of course, is how to construct this hot list. Three possible sources would be :

- Words from the document title
- Words having a very low inter document frequency
- Collocations.

In general these three sources would provide content bearing words that are specifically relevant to the article in question. Of all the suggested improvements this one would seem to be the most promising in terms of improving results.

7.4 Summary

The discussion in this Chapter was concerned with the work involved in automatically generating KB representations for queries and documents in our document retrieval application. The main steps involved were the syntactic tagging of text, removal of non-nouns, building of collocations, removal of non-content bearing terms, and the semantic tagging of the remaining terms. The amount of preparatory work involved is considerable compared to what a traditional pattern matching document retrieval system requires, refer to Section 8.4. There is scope for considerably more research into the procedures for determining content bearing terms and the correct senses of terms. However, these issues are not the central focus of this research and we believe the procedures described in this Chapter are reasonable first time approaches to these issues.

In the following Chapter we describe the operation of our retrieval engine. In Chapter 9 the results of applying our system to the task of retrieving relevant WSJ articles for a TREC query are presented.

Chapter 8 - Retrieval Engine

8. Introduction

Having generated KB representations for the documents and queries, the next step in our document retrieval system is to match these representations against each other. The discussion in this Chapter concentrates on this procedure. We also introduce the traditional information retrieval system which we use in Chapter nine as a baseline to measure the retrieval performance of our system.

Section 8.1 describes the initial design of our matching procedure. Section 8.2 presents a sample run of the system using both the information based and conceptual distance estimators of semantic similarity. The test documents introduced in Section 7.2.2 of the previous Chapter are reused in this Section. In Section 8.3 there is a recap on the design of the retrieval engine with some minor changes made based on the trial run in 8.2. The traditional pattern matching information retrieval system is described in Section 8.4 and our overall conclusions on this Chapter are presented in Section 8.5.

8.1 Design of the Retrieval Engine

The retrieval engine operates with two distinct configurations, one using the information based similarity estimator and the other using the conceptual distance approach to similarity estimation. The basic operation of the matcher is the same in

both configurations and unless stated otherwise it should be assumed that procedure applies to both configurations.

The basic querying strategy is to compare each term in the queries' KB representation, (referred to as query terms), against all the terms of each articles' KB representation, (referred to as index terms), and to aggregate all comparisons to give an overall score for the relevance of that article to the query. This unnormalised score does, however, unfairly reward articles with a large number of index terms, and therefore the score is normalised by dividing it by the number of index terms. This can be formally represented as follows, given the query KB representation :

$$Q = \langle t_1, t_2, t_3, \dots, t_r \rangle$$

and the article representation :

$$A = \langle t_1, t_2, t_3, \dots, t_n \rangle$$

the similarity between a query term and the article can be expressed as :

$$C(A, t) = \sum_{i=1}^n (Sem_sim(t_i, t) \times w_q \times w_a) \quad (1)$$

where n is the number of index terms for article A , $Sem_sim()$ is either the information based or conceptual distance³⁹ semantic similarity estimator, w_q is the weight associated with the query term t , and w_{a_i} is the weight associated with the i^{th} index term. From here the similarity of the article to the query for either configuration can be measured by :

$$Sim(A, Q) = \frac{\sum_{i=1}^r C(A, t_i)}{n} \quad (2)$$

³⁹ Given the fact that for the conceptual distance estimator, the higher the value the less similar, we inverted the value returned. This meant instead of using 0 for absolute synonymy we used a value of 0.1.

As can be seen from (1), weighting of both query and index terms is supported in this model. The weighting mechanism is, however, quite simple and does not involve user intervention. It takes advantage of the fact TREC queries and WSJ articles nearly always include titles, (as we shall see in Section 8.4 the traditional IR system does not use weights with article terms). If the query term t in (1) is a title term then w_q has a weighted value which depends on which similarity estimator we are working with, otherwise w_q is set to 1. Similarly, if the i^{th} index term from (1) is a title term from article A, w_{ai} is given a specific value, otherwise it is set to 1. From general exposure to the WSJ articles and TREC queries it was found that query titles are generally much more reliable and informative of the narrative of the query than WSJ article title words are of the articles they relate to. As such, it was decided to give a heavier weighting to query title terms as compared with article title terms. The weightings decided on for each configuration of the system are as follows :

Query & Index Title	Query Title alone	Index Title alone
1.33	1.18	1.05

Table 8-1 Term Weightings

As can be seen from the table, the values for w_q and w_{ai} are rolled into one so there is an overall weight given to the query/index term comparison.

At present the system does not take full advantage of the fact that the semantic tagger retains all ambiguity by tagging ambiguous words with all their senses, giving each sense a suitability score. Only in the case of tied rankings does the retrieval engine take more than one sense of an index term into account. As was seen in Section 7.2.1, tied rankings can only involve two senses since the tie is brought about by the locality and information content disambiguators choosing different senses as the preferred interpretation and then giving each others choice the same runner-up scores. As such, the retrieval engine checks the list of senses of an ambiguous index term for tied scores, if a tie is found both senses are included in the relevance evaluation of the article in question. This is currently accomplished by adding one of the senses as an extra index term to the article.

A further aspect of the retrieval engine is the fact that it splits long articles into a number of separate pages. A similarity score is computed for each page individually and the scores for the best page are then used to give an overall relevance evaluation for the article⁴⁰. Given the large variability in article sizes within the WSJ, see appendix C, and the fact long articles tend to discuss more than one topic, it does seem necessary to address the issue of ensuring there is no bias for or against large articles⁴¹. [Salt93] and [Hear93] report on improvements in retrieval results due to the partitioning of articles into Sections in this way. The question of course is how to decide on when to make a partition. [Hear93] talk about distinguishing between subtopic discussions and deciding on a partition break following the analysis of terms from each paragraph. Unfortunately, the RUCL parser does not tag the text with paragraph breaks, the largest syntactic division it works with is the sentence or clause. Besides which a paragraph analysis on the scale of text being dealt with here would be too time consuming. As such, we opted for a fixed page size of 150 preprocessed index terms. Given the fact that, on average, there are 125 index terms per article, (following the preprocessing stage), it would seem to make sense to try to include the majority of articles in one page. Setting the page size at 150 allows for the inevitable increase in the number of index terms due to tied rankings following the sense disambiguation. The retrieval engine for both configurations can now be expressed as follows :

$$Similarity(A, Q) = \max_{Pg_i} \left[\frac{Sim(Pg_i, Q)}{num_uniq} \right] \quad (3)$$

where Pg_i is the i^{th} page and num_uniq is the number of unique index terms in the i^{th} page. Although, page sizes are fixed at 150, pages are not all the same size. Once 150 index terms have been read from an article a page break isn't just automatically put in.

⁴⁰ It is quite possible that long documents would discuss a number of different topics. By choosing just the *best* page we felt that documents with relevant as well as non-relevant information, could be retrieved.

⁴¹ An ulterior motive for partitioning long articles can be seen in the execution time. In order to improve running time it helps considerably if all index terms involved in query/index term comparisons are loaded into memory along with all their relevant KB information.

Instead, the page break is put in following the first end of clause marker after the 150th index term. In this way the splitting across subtopics is kept to a minimum.

A final aspect of the design of our retrieval engine is the fact that we include a mechanism to restrict noise among comparisons. Our initial approach to this was to use a percentage noise threshold so only the top scoring comparisons are included in the overall evaluation of an article for a query. This effectively eliminates noise terms, both query and index, from the final score. In an attempt to prevent the exclusion of genuinely important index/query term comparisons from the overall article evaluation, any index/query term comparison involving a query title word is automatically included in the article evaluation, regardless of its value. A certain amount of experimentation is necessary to arrive at the correct value for this threshold. An initial threshold was set at 0.85, so that only the top 85% of index/query term comparisons qualify for the overall article evaluation.

In order to have a reasonable running time it was necessary to improve the code which performed the similarity estimation. In both configurations this amounted to loading into memory all the necessary information for both the query and the article. The necessary information for the information based configuration is a copy of all the synsets on the trace to the root of the HCG, using the IS-A link type. The similarity procedure thus amounted to comparing the traces of index terms against query terms to find a HCG term that subsumes both. The information content value of this node is then the estimate of the degree of similarity between both terms. If the index and query term are from different HCGs the comparison is even faster since a look up of the information content value of the root concepts of both HCGs gives the degree of similarity. Using this speeded-up code a TREC query can be run against a 1000 WSJ articles in about 5¹/₂ hours. The bulk of this time is taken up in accessing of disk based KB.

The procedure is not quite so simple for the conceptual distance configuration. For all the index and query terms, it is necessary to have the ancestral traces for all link types stored in memory. However, the problem arises with the descendant nodes.

As we demonstrated in Section 5.2.1, the conceptual distance between two nodes can be calculated by one of two ways; a common ancestral node or a common descendant node. The problem with this is that, in general, it is not possible to load all descendant nodes of a synset into memory. This is particularly so with the IS-A link type where nodes in the middle of a HCG could have thousands of descendant nodes. Thus, locating a common descendant becomes a very slow process. To deal with this we introduced a switch which allows the user to turn off the computing of conceptual distance via a common descendant for the IS-A link type. It is not expected that turning this off will greatly effect retrieval results. In order for there to be a common descendant between nodes there must exist multiple inheritance in their subgraphs and it is well known that WordNet IS-A graphs are, by and large, strictly hierarchical. This is not the case with the Part-of, Member-of, and Substance-of graphs. However, these non-hierarchical link types generally form many small, unconnected graphs, so the number of descendants to be dealt with is always quite manageable. The final running time for the conceptual distance configuration, again with a TREC query and 1000 WSJ articles, is about 19¹/₂ hours⁴².

8.2 Sample Run of the Retrieval Engine

In this Section there is a sample run of the retrieval engine. The test collection semantically tagged in Section 7.2.2 is again used here. The procedure is as outlined in the previous Section. The test query is made up of the following terms :

- Carriers
- Transport
- Lorry
- Road
- Freight Train.

⁴² Interestingly, without the switch set for calculating common descendants of the IS-A link type the running time is about 10 times slower.

The terms in bold are title terms. Looking first at the information based configuration, the results of running this query against the 'FRUIT' article could be presented as follows (the query is across the top and document terms are along the side):

	Carriers	Transport	Lorry	Road	Freight Train
Fruit	1.00	0.789	0.789	0.789	0.789
Food	0.887	0.752	0.752	0.752	0.752
Banana	0.887	0.752	0.752	0.752	0.752
Apple	0.887	0.752	0.752	0.752	0.752
Pineapple	0.887	0.752	0.752	0.752	0.752

Assuming we use all word-word comparisons, i.e. no noise threshold, this gives a total, unnormalised score of 19.73 for the Fruit article. The normalised score is arrived at by dividing this value by the number of index terms in the sample document, $19.73/5$ or 3.95. The result of comparing the query with the 'CUTLERY' article, using the information based configuration, is as follows :

	Carriers	Transport	Lorry	Road	Freight Train
Cutlery (Tableware)	1.591	1.348	1.348	0.978	1.348
Cutlery (cutter)	1.591	1.348	1.348	0.978	1.348
Fork	1.591	1.348	1.348	0.978	1.348
Knife	1.591	1.348	1.348	0.978	1.348
Spoon	1.591	1.348	1.348	0.978	1.348
Soup_spoon	1.591	1.348	1.348	0.978	1.348

This gives an unnormalised total of 39.676 and a normalised score of 6.61. Finally, the results of the third article with the information based configuration are as follows

	Carriers	Transport	Lorry	Road	Freight Train
Vehicle	3.317	2.619	2.675	1.027	2.619
Transport	3.317	2.619	2.619	1.027	2.619
Car (automobile)	2.943	2.494	3.022	0.978	2.494
Car (railway)	2.943	2.494	2.547	0.978	2.494
Sedan	2.943	2.494	3.022	0.978	2.494
Truck	2.943	2.494	3.799	0.978	2.494
Delivery Van	2.943	2.494	3.799	0.978	2.494
Cart	2.943	2.494	2.547	0.978	2.494
Articulated Lorry	2.943	2.494	3.799	0.978	2.494
Bicycle	2.943	2.494	2.547	0.978	2.494

The unnormalised total here is 120.8, with a normalised score of 12.08. The results of comparing the sample query against the sample document collection can thus be summarised as follows :

1. Transport Article - 12.08
2. Cutlery Article - 6.61
3. Fruit Article - 3.95.

These results would coincide with what would intuitively be deemed to be correct. Given the articles are only of sample size there is not much noise; nevertheless, if we introduce a noise threshold of 85%, the scores for each article change as follows :

1. Transport Article - 11.30
2. Cutlery Article - 5.78
3. Fruit Article - 3.34.

As can be seen, there is no great overall change for this case. In terms of the results themselves we see that, in general, many terms tend to get the same similarity score. Terms that are far apart conceptually, (e.g. : all terms in the query and FRUIT article), tend to get scores that are equivalent to the information content value of synsets high

in the HCG, e.g. : object (0.752), artifact (0.978), and instrumentality (1.348). This is accounted for by the fact the information based similarity estimator operates by finding the first synset in the HCG that subsumes both parameter terms. Obviously, for terms with little in common this subsuming synset tends to be near the top of the HCG. However, of particular interest from these results, is another example of the minimality property of a metric, (refer to Section 5.2.2). Two different concepts can be more similar to each other than another concept is to itself, (refer to Section 5.1.2). The query term 'transport' is from the same synset as the 'transport' index term in the 'Transport' article but their similarity is only 2.619, the information content value of the synset they're both from. This is the same score as was given to the similarity between the terms 'Transport' and 'Freight Train', (thus 'Transport' must be a direct ancestor of 'Freight Train'). The effect of this can be more clearly seen below :

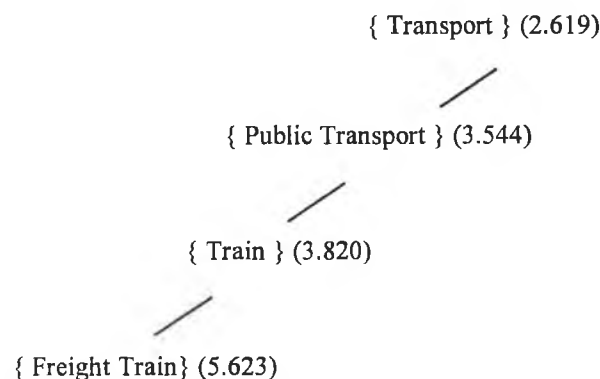


Figure 8-1 Non-metric Information based Similarity Estimator

above is an extract from the KB showing the relationship between the concepts *Freight Train* and *Transport*. The numbers in brackets after the synsets are the information content values. From here we can see the information based estimate of the similarity between *Transport* and *Freight Train*, (2.619), is the same as the estimated similarity between *Transport* and itself, (2.619). This is bound to pose certain difficulties in the use of the information based estimate of similarity in the retrieval engine. However, the degree of difficulty posed is not easily measured and

certainly the information based configuration of the system in the sample run did produce promising results.

The same query was subsequently run against the sample document collection using the conceptual distance similarity estimator in place of the information based one. The results from the three articles are tabulated as follows :

	Carriers	Transport	Lorry	Road	Freight Train
Fruit	0.124626	0.098396	0.073368	0.101471	0.082257
Food	0.103563	0.088028	0.066627	0.090613	0.0743
Banana	0.149031	0.126678	0.086633	0.1321	0.10007
Apple	0.102923	0.087481	0.066313	0.090033	0.07391
Pineapple	0.104199	0.088574	0.066939	0.091191	0.074688

This gives an unnormalised total of 2.344 for the 'Fruit' article. Normalising, again by dividing by the number of index terms, gives an overall score of 0.4688 for this article and the sample query. The results for the 'Cutlery' article are :

	Carriers	Transport	Lorry	Road	Freight Train
Cutlery (Tableware)	0.250125	0.212585	0.119717	0.137212	0.146994
Cutlery (cutter)	0.252653	0.214731	0.120395	0.138102	0.148017
Fork	0.224065	0.190476	0.112372	0.127649	0.136073
Knife	0.223714	0.19015	0.112259	0.127502	0.135906
Spoon	0.225785	0.191939	0.11288	0.128304	0.136818
Soup_spoon	0.192419	0.163559	0.102428	0.114969	0.121758

The unnormalised score here is 4.811, this is normalised to give a score of 0.801 for the 'Cutlery' article. Finally, results for the 'Transport' article are :

	Carriers	Transport	Lorry	Road	Freight Train
Vehicle	1.35318	1.068376	0.395101	0.184434	0.341297
Transport	13.33	13.33	0.288434	0.222916	0.501505
Car (automobile)	0.439174	0.373274	0.379075	0.135117	0.209293
Car (railway)	0.457247	0.388651	0.235183	0.13708	0.214041
Sedan	0.36062	0.30656	0.310463	0.125251	0.186532
Truck	0.458926	0.390168	0.92081	0.137268	0.2145
Delivery Van	0.327118	0.278009	0.471698	0.120207	0.175562
Cart	0.457457	0.388802	0.235239	0.137099	0.214087
Articulated Lorry	0.378788	0.321958	0.613874	0.127747	0.192123
Bicycle	0.371609	0.315856	0.206398	0.126775	0.189934

The unnormalised value for the 'Transport' article is 43.048, this converts into a normalised score of 4.304 for the article as a whole.

The results for the conceptual distance configuration can thus be summarised as :

1. Transport Article - 4.304
2. Cutlery Article - 0.801
3. Fruit Article - 0.468.

We see the results are again satisfactory and in agreement with both the information based configuration and what would be intuitively thought of as being correct. If a noise threshold of 85% is used with the conceptual distance configuration, the results change as follows :

1. Transport Article - 4.213
2. Cutlery Article - 0.728
3. Fruit Article - 0.428.

Once again there is no great overall change, explained because of the lack of noise due to the small size of the sample.

The results for the conceptual distance configuration differ greatly from those of the information based approach, even in terms of the fact that no two terms tend to get the same similarity score. However, the final interpretation of the results from both configuration yields the same overall result, and this very satisfying. Of particular interest from these results is the weakness apparent with the conceptual distance similarity estimator. As with the information based approach, the conceptual distance configuration is not without fault. If we look, for instance, at the similarity between 'cutlery' and 'transport', (4.657), we see according to this configuration these two concepts have more in common than, 'freight train' and 'car (as in automobile)', (4.778). If we examine why this is so we see from Figure 8-2 that the general structure of the HCG is in fact to blame :

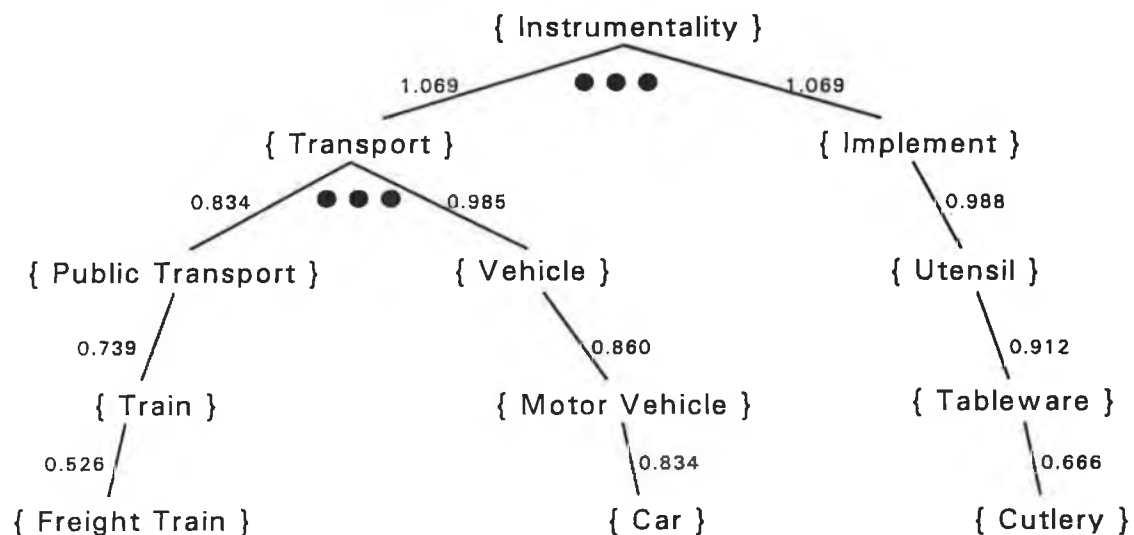


Figure 8-2 Weighted KB Extract

Due to the fact that the 'transport' synset is so high up in the HCG, most concepts in the middle to high Sections of the HCG would be "geographically" close to it and would therefore be deemed to be conceptually similar to it. This is the case with the value returned for the similarity between 'cutlery' and 'transport'. 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 KB.

8.3 Conclusions on Retrieval Engine

Our main concern with the application of the retrieval engine in the large scale was the level of noise it has to deal with. With an average of 125 preprocessed index terms per article, a lot of these terms could be noise terms. Both the information based and conceptual distance similarity estimators are, in general, quite good at estimating the similarity between terms that are relatively similar, 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 similarity between a banana and an apple and a banana and a car, however, the similarity values become more unclear if we are 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. The initial design of our model, as described in the previous Section, does not reflect this. We believe the overall effect will be seen in how noise is dealt with by the system. Certain documents may have a number of particularly relevant terms that will receive good similarity scores, however, these terms will be overshadowed by noise terms receiving bad similarity scores. Traditional pattern matching retrieval systems deal well with noise; if there is no direct match then the term is simply ignored. The downside of this is, of course, that relevant documents are not retrieved by the overlooking of terms with approximately similar meanings and non-relevant documents are retrieved because of the presence of homonyms. From here it would seem that the best results would be achievable if our system extended its current handling of noise terms to include some of the features of noise handling in traditional IR systems. This could be accomplished by replacing the noise thresholding suggested in Section 8.1 by a mechanism which would disregard query term/index term matches that were outside a certain threshold⁴³. If we were comparing two

⁴³ The values for this threshold are set in the Chapter 9 following a number of experiments.

dissimilar terms the comparison would be omitted from the overall evaluation for that article and query. The normalising factor in (3) changes to reflect the decrease in the number of index terms being dealt with :

$$Similarity(A, Q) = \max_{Pg} \left[\frac{Sim(Pg, Q) \times terms_used}{num_uniq} \right] \quad (4)$$

where terms used refers to the number index/query term comparisons that are above the noise threshold, and all other variables are as per (3).

This procedure would hopefully allow the inclusion of approximately similar terms, (e.g. articulated_lorry and truck), thus including relevant documents indexed under different although semantically similar index terms. The ability of the system to exclude documents indexed with homonyms of the query terms will be determined by the competence of the sense disambiguator in the choosing of correct senses. This could thus provide a rough test bed for evaluation of our semantic disambiguator.

8.4 Baseline Retrieval System

In this Section we describe the retrieval system whose performance we use in Chapter nine as a baseline in the evaluation of our semantic document retrieval system. This system is an example of a traditional pattern matching information retrieval system which uses term frequency weights to weight the importance of index/query term comparisons⁴⁴.

The operation of the baseline retrieval system is very different from that of our semantic retrieval system. From Chapter 7 it can be recalled how the text of queries

⁴⁴ We would like to thank Fergus Kellely, the developer of this IR system.

and WSJ articles had to go through a number of preprocessing steps. This is not necessary with the baseline retrieval system. The only preprocessing of the text necessary is the stripping of punctuation marks and stop words, and the stemming of remaining words. Stop words in this situation are conjunctions and words like 'the', 'a', 'it', etc.. With only stop words being removed index and query terms can be made up of nouns, verbs, adjectives, and adverbs. This can lead to a big increase in the number of noise terms. We will see later how this noise is handled. The stemming procedure reduces words to their base forms. The stemming algorithm used is a variation of Porter's stemming algorithm, [Port80]⁴⁵. This differs from the morphological procedure used in our retrieval system insofar as the base form of the word might not actually be a real word, for instance, the word *computers* is reduced to *comput*.

The retrieval procedure starts off by searching all documents for the query terms. The comparison process is pattern matching so the query terms are matched character by character against the index terms of all articles. This can be a lengthy process so to reduce the amount of time taken, an inverted file is used. The inverted file can be thought of as a large table of terms. This table has a row for each unique term in the collection and the first column contains an entry for the number of occurrences of this term in the collection, (known as the out-document frequency), the second column contains a count of the number of articles the term appears in, and for each of these articles the remaining columns contain an entry for the article ID and a count of the number of occurrences of the term in the article, (known as the in-document frequency). The article ID and in-document frequency are collectively referred to as a posting. With this arrangement the retrieval strategy amounts to locating query terms in the inverted file. The rows corresponding to the query terms are then extracted to form a table for that query. The rows of this table are then sorted in order of the length of their posting lists with the term with the least number of postings coming first. The postings themselves are sorted on in-document frequency with the article having the most number of occurrences of the query term appearing first. This gives us something like the following :

⁴⁵ The only real difference can be seen in the size of the stop list used.

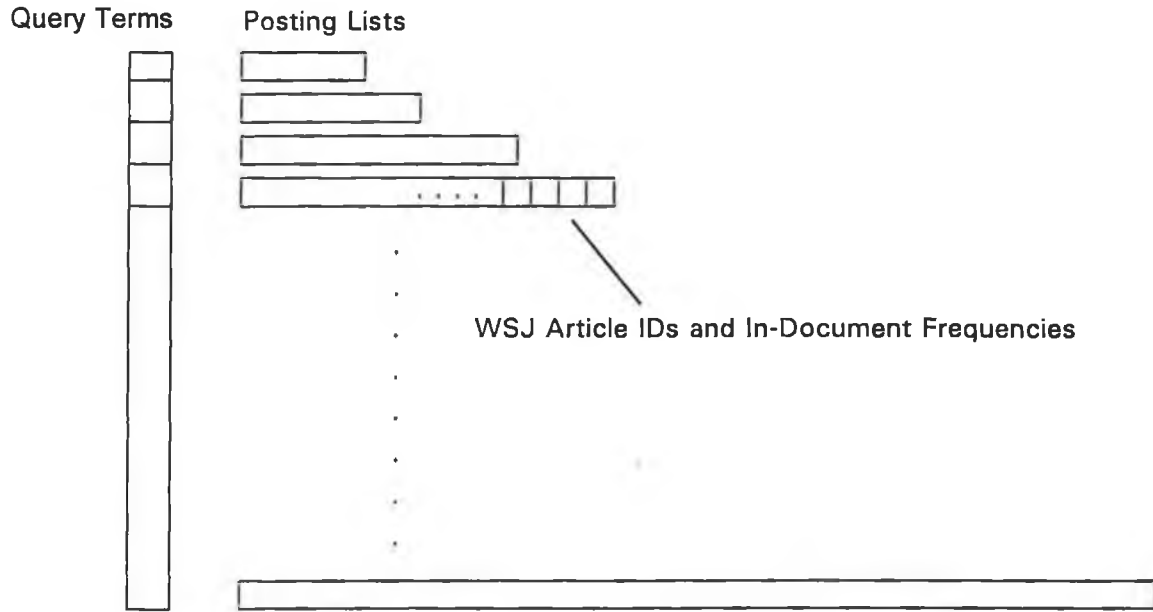


Figure 8-3 Structure Generated for Queries

As well as this structure there is a large matrix in which there is a row for each WSJ file and columns represent particular articles within a WSJ file. This matrix is used to record the relevance scores of articles for queries. The process starts by reading the posting list of the first query term and for each article ID in this posting list the relevance score for that article is incremented in the matrix by :

$$Rel_weight * tf_IDF_WEIGHT \quad (5)$$

where Rel_weight is calculated by assigning the first query term an arbitrary weight, AW , and decreasing this weight for each term in turn by :

$$\frac{AW}{2 * Tot_query_terms} \quad (6).$$

The $tf*IDF$ weight, [vanR79], is as follows :

$$In_freq * Log\left(\frac{NP}{Tot_doc}\right) \quad (7)$$

where *In_freq* is the in-document frequency, *NP* is the number of postings for that query term, and *Tot_doc* is the total number of articles. It is possible to assign weights to query terms depending on whether they are query title terms or not. This changes (4) to give :

$$Qterm_weight * Rel_weight * tf_IDF_weight \quad (8)$$

Only the full posting list of the first query term is fully processed. As we descend through the query terms the percentage of the posting list processed is decreased. Finally, after a certain threshold of articles⁴⁶ have been attributed a score from this procedure, all new articles in subsequent postings are ignored. Effectively, the remaining postings are only used to increment the relevance scores of this threshold set of articles. These features of the *tf*IDF* system are peculiar to our system and were intended to maximise retrieval performance.

The performance of this system will be presented in the following Chapter, however, in terms of the speed with which it operates we have found that on average it processes a TREC query against 550 Mbytes of text in approximately 30 seconds. This is orders of magnitude faster than our retrieval system which when operating with the information based configuration, processes an average query in approximately 5 and a half hours and when operating with the conceptual distance similarity estimator takes nearly 20 hours per query. Therefore regardless of retrieval performance, further research will have to be carried out into improving the speed of our system. The most obvious approach is to have the KB in memory at all times as this would effect a great improvement in speed.

⁴⁶ For the WSJ collection this was set to 6,800

8.5 Summary

In this Chapter we described how queries and documents are matched in our semantic knowledge based document retrieval system. A small sample run, using the same test collection used in the demonstration of the semantic disambiguator in Chapter 7, was presented for both the information theoretic and conceptual distance configurations. In Section 8.4 we described the operation of a traditional information retrieval system. Both systems differed considerably in terms of how they worked, the speed with which they performed, and the amount of preprocessing of the document and query text required. The retrieval performance of both systems is compared in the following Chapter.

The TREC evaluation procedures are used to compare the performance of our semantic retrieval system against the pattern matching approach. A description of this evaluation mechanism along with a description of the experimental design is to be found in the next Chapter. The results of a number of experiments to set values for the noise threshold parameters described in Section 8.1 and 8.3 are also presented in Chapter 9.

Chapter 9 - Evaluation

9. Introduction

In the previous two Chapters we have described the operation of our semantic knowledge based system in a document retrieval application. In this Chapter we look at the results of an evaluation of this application. Up to now we have only described TREC as a large text corpus with a set of queries and relevance assessments for these queries. In Section 9.1 of this Chapter we give a complete description of TREC and its automatic evaluation mechanism.

Before carrying out a full scale test of our system, however, we must first decide on optimum noise threshold values for both our information based and conceptual distance configurations. Details of how this is performed can be found in Section 9.2. The performance of these configurations are then compared against the performance of a traditional pattern matching information retrieval system. Details of this comparison can be found in Section 9.3. Finally, a discussion of these results along with conclusions are presented in Section 9.4.

9.1 TREC

The Text REtrieval Conference (TREC) was created with the related goals of :

- (a) Providing a large text corpus as a test bed for evaluating IR techniques.
- (b) Providing an environment within which different approaches to retrieving information could be compared.

The first point is made clearer by the fact that prior to TREC most prototype IR systems were tested on small collections, generally measured in Megabytes of text, whereas most commercial applications involved Gigabytes of text. This imbalance in size was often thought to lead to improper conclusions being drawn from IR research. The issue of the performance of research prototypes in the large scale was always an issue of contention. The second point regarding the comparison of approaches to IR has also been a long recognised stumbling block to progress in IR :

“Much of the work in IR has suffered from the difficulty of comparing retrieval results. Experiments have been done with a large variety of document collections, and rarely has the same document collection been used in quite the same form in more than one piece of work. Therefore one is always left with the suspicion that worker A’s results may be data specific and that were he to test them on worker B’s data they would not hold.”, [vanR81].

TREC’s goal was to overcome these problems. In the following subsections we describe the TREC corpus and TREC’s automatic evaluation mechanism.

9.1.1 TREC Corpus

The TREC corpus was taken from the results of the TIPSTER project and can be separated into the following parts :

1. The Documents
2. The Queries
3. The Relevance Assessments

The document texts are made up of a number of newspaper, US government and newswire sources. Following is the list of sources used in the full 3 Gigabytes of text

Source	From	To	Size (Mbytes)
Wall Street Journal	1986	1992	550
AP Newswire	1988	1989	514
Information from Computer Select Disks	-	-	439
Federal Register (US)	1988	1989	469
Short Abstracts from the Dept. of Energy (US)	-	-	190

Table 9-1 Source material for TREC document Collection

As mentioned in Chapter 7, we at Dublin City University are using only the Wall Street Journal text collection⁴⁷. Refer to appendix I and Section 7.1 for a detailed description of this collection. The TREC queries are also described in Section 7.1 and sample queries can be found in appendix I.

In order to evaluate an IR system it is first necessary to know what the relevant documents for individual queries are. This set of relevant documents along with the set of queries they pertain to are known as the relevance assessments of a text corpus. The most obvious way of constructing relevance assessments is to manually read every document in the collection to determine its relevance to each query. However, with up to 742,611 documents in the TREC collection, and a set of 50 queries, such a procedure involves 37,130,550 query-document comparisons. Obviously this is not feasible, so a less rigorous method was needed.

This less rigorous approach is based on an earlier attempt to resolve the same problem for the British Library, [Spar75]. This method is known as *pooling*, and, as the name suggests, involves the combining of the results of using a wide variety of IR systems to determine relevant documents for a query. The steps are as follows :

⁴⁷ Participants of TREC are divided into one of three categories and category B participants only use the 550 Megabytes of the Wall Street Journal.

1. Divide the results from each system into results for each query.
2. For each query, select the top 1000 ranked documents for input to the pool, as ranked by as many different IR approaches as possible.
3. For each query, merge the results.
4. For each query, remove duplicate documents.

This pooled set of documents is then manually assessed by a bank of readers and documents are discarded if they are considered non-relevant to the query.

9.1.2 TREC Evaluation Mechanism

As stated above one of the main aims of TREC is to provide an unbiased environment for comparing the performance of many different IR systems. To facilitate this, TREC organise a controlled evaluation of IR systems on an annual basis. At a set date 50 queries without relevance assessments are distributed to all participants of TREC. Around two weeks is then allowed for all IR groups to compute the top 1000 retrieved documents for each query and return these ranked lists to TREC. The top 100 documents from each returned rank list were then used to form one large pool of documents. Duplicates were then eliminated. The results of this procedure in TREC-2 are illustrated below :

TREC-2	Maximum	Actual
Unique Documents per Topic	4000	1106

Table 9-2 Overlap of Submitted Results

Out of the maximum of 4000 possible unique documents (40 runs times 100 documents), over one quarter of the documents were found to be unique. This lack of overlap indicates the degree of heterogeneity of IR systems used in TREC. However, other reasons for the lack of overlap could include the very large number of documents that contain the same terms as the relevant documents as well as the very different sets of terms in the constructed queries.

The unique documents from the pool are subsequently manually evaluated and the documents remaining after this evaluation make up the relevance assessments. These are then sent out to the TREC participants who use them to evaluate their individual approaches.

The evaluation mechanism is based on precision and recall. These two values have long been used by the IR community to measure the effectiveness of retrieval systems :

“It is recall and precision that attempt to measure what is known as the effectiveness of the retrieval system. In other words it is a measure of the ability of the system to retrieve relevant documents while at the same time holding back non-relevant ones.”, [VanR81].

Precision is defined as the proportion of retrieved documents that is relevant, while recall is the proportion of relevant material retrieved. Precision and recall values for a retrieval system tend to be inversely related insofar as at high levels of precision, recall levels tend to be quite low, and at high recall levels, precision values tend to be low. This can be made clearer by example. If there are 20 documents retrieved for a query as follows :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
*	*	*	*	*		*		*		*				*			*		*

Table 9-3 Ranked List of 20 retrieved documents

An asterisk indicates a document are relevant. If we take precision and recall values after 5, 10, 15 and 20 documents are retrieved, we get the following values :

At rank position	Precision = $\frac{\text{Total relevant retrieved}}{\text{Total retrieved}}$	Recall = $\frac{\text{Total relevant retrieved}}{\text{Total relevant}}$
5	$\frac{5}{5}$ (1.00)	$\frac{5}{11}$ (0.45)
10	$\frac{7}{10}$ (0.70)	$\frac{7}{11}$ (0.63)
15	$\frac{9}{15}$ (0.60)	$\frac{9}{11}$ (0.82)
20	$\frac{11}{20}$ (0.55)	$\frac{11}{11}$ (1.00)

Table 9-4 Sample Precision and Recall Values

Taking precision and recall values at different rankings in this manner allows the derivation of precision and recall graphs.

The organisers of TREC have developed an evaluation tool that accepts a ranked list of retrieved documents and produces standard precision and recall figures. The application calculates averaged precision values for standard recall points in the range 0 to 1.0, (in steps of 0.1). It is possible to enter a ranked list for one query or more queries. In the case of multiple queries, a single graph is produced which represents the interpolated performance of the retrieval system across a number of queries. By using these graphs, it is possible to compare the performance of one retrieval system against another. Refer to Figure 9-1 for a sample graph comparing two systems. System A has a higher precision at the low recall end of the graph and therefore is more accurate. However, system B has higher precision at the high recall end of the graph and therefore will give a more complete set of relevant documents, assuming the user is willing to look further down the ranked list, (however, with a 1000 documents in TREC, this may seem impractical).

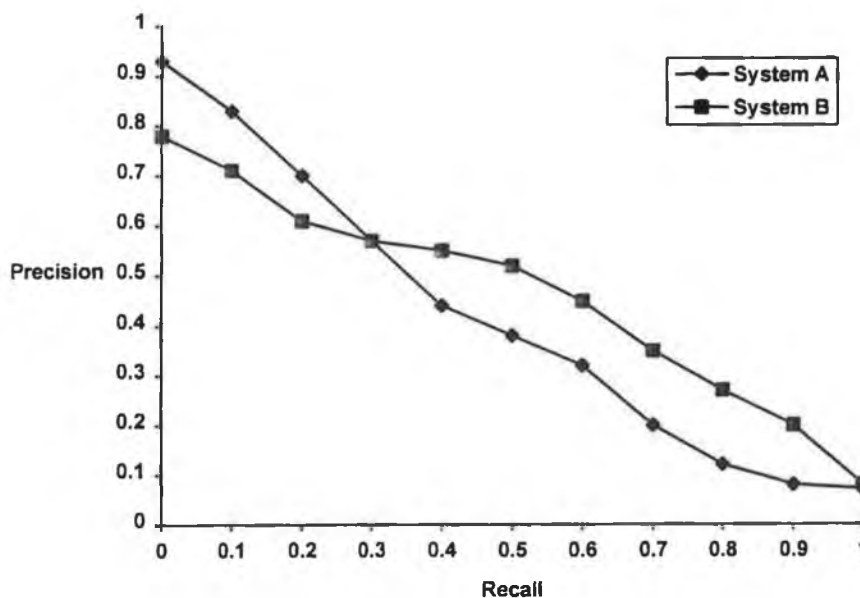


Figure 9-1 Sample Precision/Recall graph Comparing two Systems

9.2 Optimum Noise Threshold

In this Section we will describe the set of experiments carried out to determine the optimum noise threshold values for both the conceptual distance and information based configurations of our system. It was decided in Section 8.3 of the previous Chapter that absolute value noise thresholds should be used in place of a percentage threshold. The basic reasoning had to do with the fact that a percentage threshold assumes there is some overlap between an article and a query. However, in many cases the query and article will have nothing in common, yet the percentage threshold still takes a top fixed percentage of comparisons. In contrast, an absolute value threshold operates on an individual index-query term match. If a match is above a certain threshold, we can be sure the index and query terms are related thus highlighting the possibility of the article being relevant. Otherwise, the match is ignored and in cases where an article has absolutely nothing in common with a query, all index-query term matches are ignored, effectively eliminating the article from the set of articles considered relevant.

Of course the question still remains, what values are assigned to these thresholds? Clearly the same value cannot be assigned to both the conceptual distance and information based configurations. Each configurations approach to estimating semantic similarity differs greatly and this is reflected in the different range and orientation of values they return. As such, a separate set of experiments is needed for each configuration. The actual value of the threshold determines both the extent of noise allowed and the degree to which the system emulates a pattern matching system. A noise threshold value of 0 for the conceptual distance configuration and a value of approximately 7.7 for the information based configuration effectively makes our system a pattern matching system with a special handling for homonyms. As the noise

threshold gets bigger, (or smaller for the information based configuration), the more flexible or lenient our interpretation of relatedness becomes.

The remainder of this Section is organised as follows. In Section 9.2.1 we discuss the design of these experiments. Results of the experiments are presented in Section 9.2.2 and finally, Section 9.2.3 contains a general discussion the on outcome of the experiments.

9.2.1 Experimental Design

The obvious approach to arriving at an optimum configuration of our system in terms of noise thresholding is to use a training set of TREC queries on the WSJ corpus. The restraining factor in these experiments is the computational time required. This point was mentioned in Section 8.1, however, we will elaborate on it here. If we examine the operation of the retrieval engine in terms of disk accesses and in-memory operations we can break it down as follows :

Assuming an average of 15 query terms per query, 125 index terms per document, average depth of a synset as 5, and the average number of descendants for a synset could conservatively be estimated to be 5 also. For the information based system, access is required to all index/query term synsets as well as their ancestor synsets. This gives us :

$$(5 * 15) + (125 * 5 * 1000) = 625,075 \text{ disk accesses}$$

For the conceptual distance system access is needed to both descendant and ancestor nodes of index/query term synsets. The number of accesses is also escalated by the fact the conceptual distance uses the part-of, member-of and substance-of relational links, as well as the is-a links. However, ignoring this, we have :

$$(15 * 5 * 5) + (125 * 5 * 5 * 1000) = 3,125,375 \text{ disk accesses.}$$

For both systems there is a pairwise comparison of all index terms against query terms for 1,000 documents, this amounts to :

$$(15 * 125 * 1000) = 1,875,000 \text{ comparison operations.}$$

Comparison operations involve tracing through link list structures, backtracking because of tangled hierarchies caused by multiple inheritance, and at each stage performing arithmetic or comparative floating point operations. Again the comparison operation is considerably more complex in the case of the conceptual distance, refer to Section 5.2.1

A similar investigation of the sense disambiguator is as follows :

Again assuming there are 125 index terms per document, we would have a minimum of $125 * 1000 = 125,000$, disk accesses if there was no ambiguity. Every time an ambiguous term is encountered the four sense disambiguators are applied to the 5 context words on either side of the ambiguous word. For the locality disambiguator we again need access to the ancestor synsets, refer to Section 7.2.2. As such, with an average of 75% of the terms ambiguous, (refer to Section 7.2.1), and an average synset depth of 5, the disk accesses involved in the sense disambiguator can be estimated to be :

$$125,000 + ((0.75 * 125,000) * 5 * 11) = 5,281,250.$$

The in-memory operations involved for the sense disambiguator are quite difficult to summarise, but we can say this operation is required to determine whether a term is ambiguous and for ambiguous terms there is at least one operation carried out by each sense disambiguator :

$$125,000 + ((0.75 * 125,000) * 4) = 500,000$$

From these break downs it is not difficult to see why the retrieval engine takes up to 24 hours per query and the sense disambiguator takes a similar length of time to disambiguate 1000 documents. In this research we did not attempt engineer a faster system, our interest was primarily to build a flexible prototype system which allowed us to vary as many parameters as we could.

The other factors in these experiments are the number of queries to be used and the range of threshold values to be tested. In order for the results of the experiments to be non-biased it is necessary to use as many queries as is possible. Experiments using too few queries would be biased toward the subject matter of those queries. Also, in order to obtain the most optimum noise threshold value it is necessary to repeat the experiments a large number of times, changing the noise threshold values only slightly on each iteration. There is an obvious trade-off here between optimality and impartiality.

In our approach we opted for impartiality at the possible expense of an optimal noise threshold. A set of 12 queries were used with four noise threshold values for both the information based and conceptual distance configurations. This resulted in 12x4x2 iterations of the retrieval engine as well as 12 iterations of the sense

disambiguator. This still entailed a large amount of computational processing time. Considering these experiments were only to set threshold values, and the main system testing was still to come, we had to consider other methods to reduce the amount of time involved. One obvious option was to reduce the size of the document set for each query. If the number of articles searched through for each query was reduced from 1000 to say 200, the amount of time needed to carry out the experiments could be cut down to a more reasonable time scale. The value of 200 is not an unreasonable figure. In the first TREC, TREC-1, participants returned just the top 200 retrieved articles and the decision to extend the number of retrieved articles to 1000 for TREC-2 and TREC-3 was made because very often the top 200 retrieved articles included very few relevant articles. This problem was overcome in our situation by ensuring there was always a certain percentage of relevant articles amongst the 200 in the test set of each query. The actual process of selecting the set of 200 articles for a query was as follows :

- (a) Take the top 100 articles retrieved by a pattern matching system
for the query in question
- (b) Augment this 100 articles by as many relevant articles as is
necessary to ensure there are at least 40 relevant articles present
- (c) Make up the remaining article from randomly selected articles,
ensuring none of these articles were already selected by (a) or ((b).

There was no reason why we chose the queries we did for the experiment, (beyond the fact that the syntactically parsed text for certain query results was present while that for others was not). As can be seen from appendix H, the queries vary greatly in terms of subject matter, thus minimising the amount of subject bias. The text of the TREC queries were sent through the same preprocessing steps as the text of the documents, however, the resulting parsed query text was checked for correctness and manually altered in certain situations. This is in line with what is allowed in TREC; as long as the sense of the query is unchanged, it is perfectly legal to alter the terms making up a query. Most changes involved the introduction of collocations not

found by the collocation builder or the replacement of particularly polysemous query terms by equivalent alternatives.

9.2.2 Experimental Results for Noise Thresholds

The results of our experiments to determine the optimum noise thresholds for the two configurations of our systems are presented in this Section. The results for each configuration are displayed in the form of a precision/recall graph giving the performance of the configuration at a given threshold interpolated averaged over 12 interpolated queries.

In arriving at the set of threshold values to be tested for each configuration we were attempting to capture a peak performance. As was outlined in the previous Section, given the time constraints involved, an optimum threshold was not being sought. Instead we were simply looking for a situation where given three threshold values A, B, and C, where $A < B$ and $B < C$, the performances of the system using A and C noise thresholds are less than the performance using B. In order to arrive at values for A, B, and C, a number of informal experiments were carried out with values which were intuitively thought to be appropriate. As a result of these experiments, the following threshold values were selected for the full runs :

Conceptual Distance	Information Based
7	1.2
9	1.3
11	1.4

Table 9-5 Initial Absolute Noise Threshold Values

In order to have a baseline to compare the performance of both systems for each of these noise thresholds, we included an experimental run with effectively no noise threshold. In this run any query term/index term match which involved terms from the same HCG were included in the evaluation of the relevance of the document for the

query. Strictly speaking this is not the same as no noise threshold at all, but rather a noise threshold of approximately 0.55 for the information based configuration, and a threshold of approximately 20 for the conceptual distance threshold⁴⁸.

The results of using these noise threshold values are presented in figure 9.2 for the conceptual distance configuration and figure 9.3 for the information based configuration. The average precisions for both configurations at each threshold value is shown in Table 9.6.

Configuration	No Threshold	1.2 / 7	1.3 / 9	1.4 / 11
Conceptual Distance	0.0819	0.1639	0.1697	0.1701
Information Based	0.1538	0.1537	0.1634	0.1538

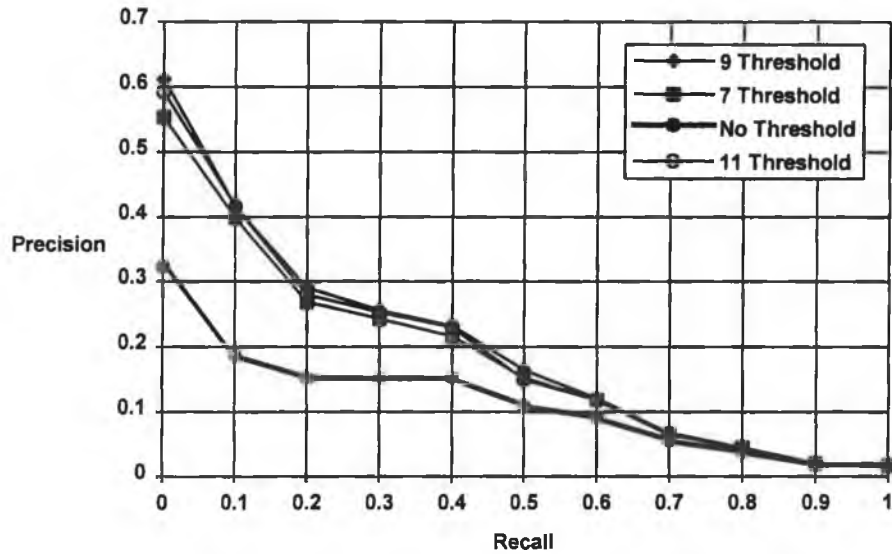
Table 9-6 Average Precisions for each configuration

In both configurations, the performance at *No Threshold* is considerably poorer at all levels of recall. This result supports our hypothesis that the relative accuracy of judgements of similarity is dependent on how related the two terms being judged are, (see Section 8.3). As can be seen from figures 9.2 and 9.3 the performance of both configurations is improved by including only similarity comparisons between related terms in the determination of relevance of a document for a query. It appears from Figure 9-2 and Figure 9-3 that the information based configuration is less sensitive to noise than the conceptual distance configuration. This is probably explained by the method of using absolute noise thresholds to deal with noise. Values returned as estimates of similarity by the conceptual distance configuration are more directly comparable than those returned by the information based configuration. For instance, a similarity value of 0 for the conceptual distance configuration always means the terms being compared are synonymous. In contrast, the information based value of similarity for synonymous concepts depends on the pair of concepts being

⁴⁸ The information content of the root synset { Entity } is 0.549, and a conservative estimate of the average conceptual distance between a node and its root node is approximately 10.

compared⁴⁹. In Figure 9-2 it can be seen that the optimum noise threshold value for the conceptual distance configuration was captured between the values of 7 and 11.

(A)



(B)

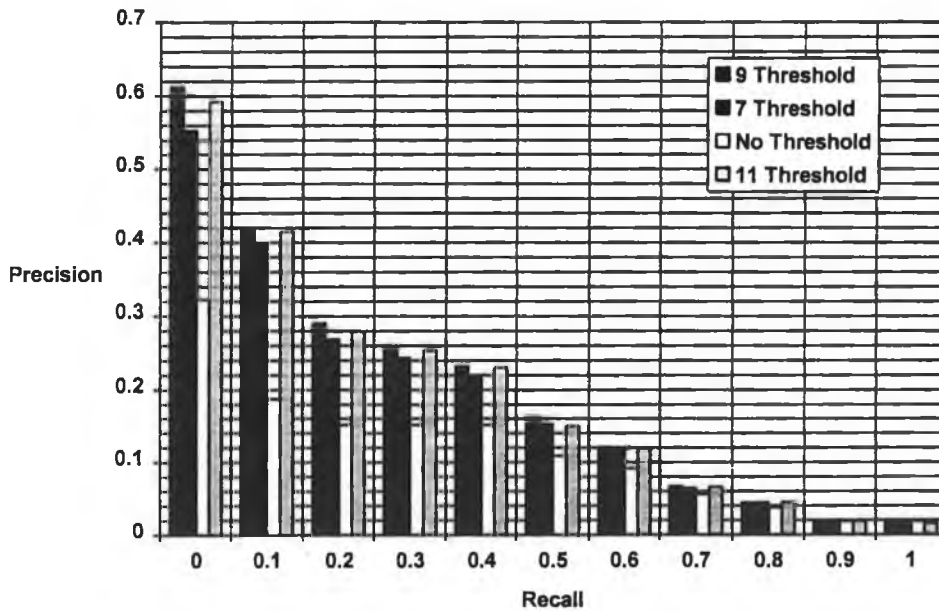
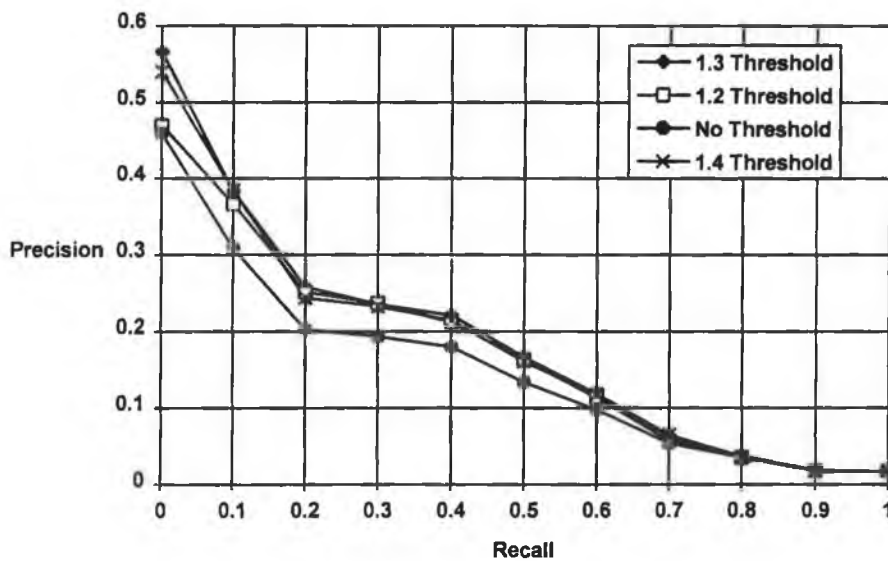


Figure 9-2 Optimum Noise Threshold For Conceptual Distance Configuration

⁴⁹ This relates back to the information based estimator's breach of the minimality property of a metric, (see Section 5.2.2 and Section 8.2).

The performance at the value 7 is clearly worse than at the value 9, and although not as clear, the performance at the threshold of 11 is slightly worse than that at 9. This is probably seen more clearly with the bar chart display of Figure 9-2.

(A)



(B)

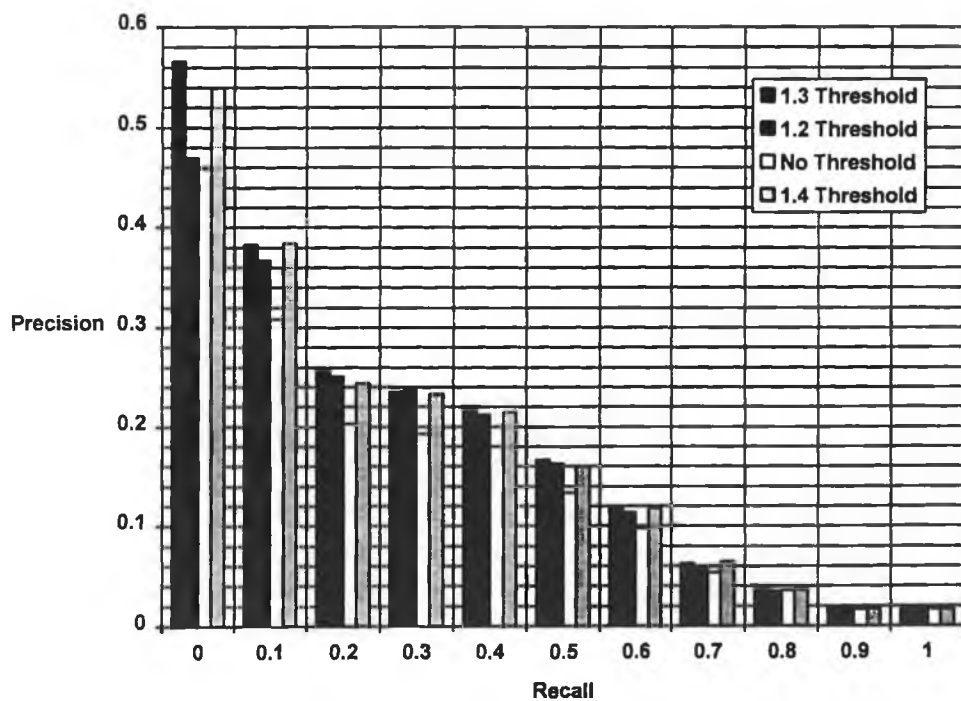


Figure 9-3 Optimum Noise Threshold for the Information Based Configuration

In the case of the information based configuration, the optimum noise threshold was found to lie between the values of 1.2 and 1.4. Again this can be more clearly seen with the bar chart display of Figure 9-3 (b).

As a result of these experiments an absolute noise threshold value of 1.3 was chosen for the information based configuration and a value of 9 was selected for the conceptual distance configuration.

9.2.3 Discussion of Results

Overall the results for both configurations were promising. The possibility for error in our application is enormous. Errors could be generated at each of the following stages :

- Syntactic tagging
- Text preprocessing
- Semantic disambiguation
- Query matching.

Had the error level at each stage been substantial, all errors would have aggregated up to produce very poor results. Beyond this it has to be born in mind that the only variable which we attempted to fine tune was the noise threshold. Very many of the other variables were set at values that were simply considered to be intuitively correct. A brief listing of the more important variables includes :

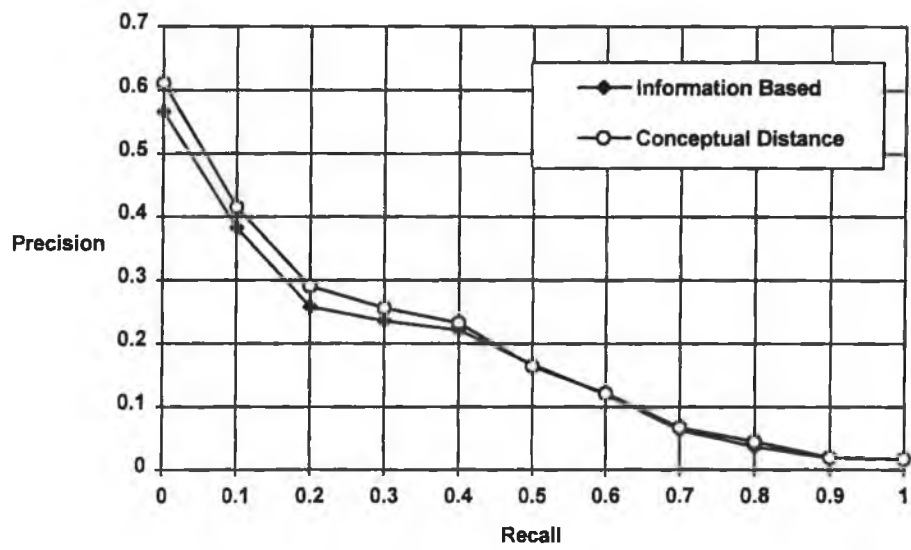
- Weighting of links
 - Values for Min and Max in the density normaliser
 - Weighting of surrounding fanouts in the local density estimator
 - Depth scaling : combination of information content values and HCG depth

- Text Preprocessing
 - Document frequency threshold for inclusion in the exception list
- Sense Disambiguation
 - Window size
 - Relative power of each disambiguator
 - Set of weightings for the locality and information content disambiguator
 - Relative power of the information content disambiguator in the case where the ambiguous word is dominant in a synset as opposed to when it is not
- Querying
 - Page size
 - Noise threshold
 - Weighting of query and index title terms
 - Setting synonymy for conceptual distance estimator at 0.1.

Many of these variables are inter-connected insofar as a change to the value of one effects the value of another. This particularly high number of dependent variables effectively rules out the possibility of an evaluation of an optimised configuration of our system. However, as the results prove, the configurations we are operating with are not a bad starting point.

Due to the fact the same queries and document sets were used in the determination of optimal noise threshold values for both the conceptual distance and information based configuration's, we were presented with an opportunity to directly compare each configurations performance. The average precision for both configurations can be seen in Table 9.5 and Figure 9-4 shows the precision recall graph. As can be seen, both configurations have comparative performances, with the conceptual distance configuration performing marginally better than the information based configuration. Reasons for this are difficult to determine but we believe the violation of the metric property of minimality is a primary factor in the lower performance of the information based configuration.

(A)



(B)

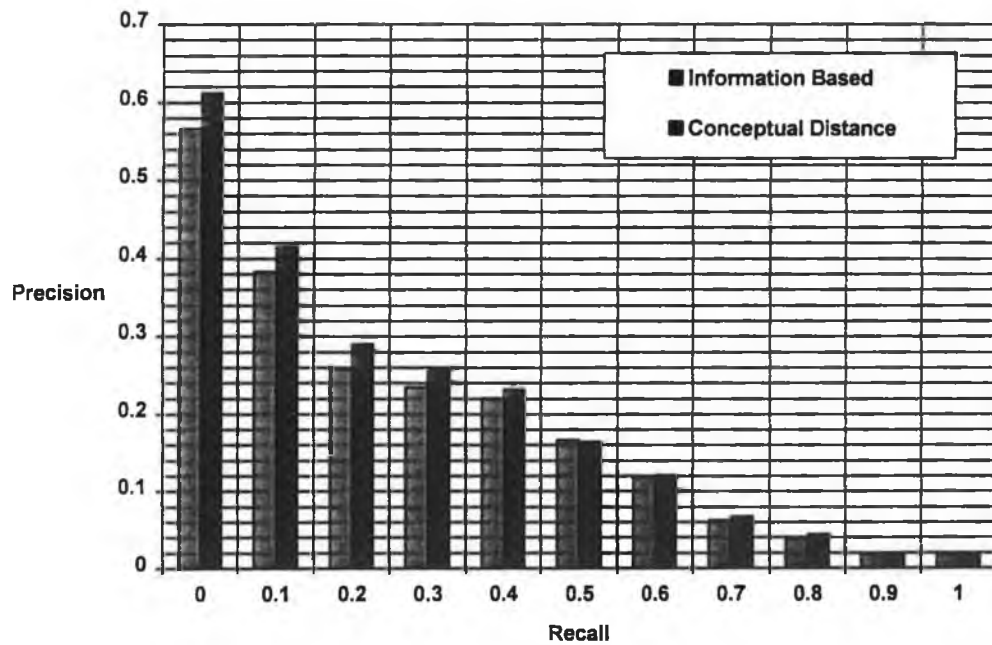


Figure 9-4 Information Based Vs. Conceptual Distance System

Appendix I contains a break down of results for each query. These results help determine whether the better average performance for the conceptual distance system results from better performance on most of the test queries or from comparable performance on most queries and significantly better performance on other queries. Also, by producing results for individual queries we answered questions concerning our representation of the TREC queries. If both systems performed badly on a specific query then we could postulate that the reason for the poor performance is related to our KB representation of the query. Upon examination of the precision recall graphs for each query we see, quite surprisingly, that in certain queries the information based system outperformed the conceptual distance system. For test queries 1, 2, 6, and 8 the information based performed best and in queries 5, 7, and 12 the performance of both systems were comparable. The conceptual distance system performed very poorly on queries 1 and 8, and although the conceptual distance system outperformed the information based system in queries 3, 4, 9, 10 and 11, the performance of the information based system was not very poor on any of these queries. This would suggest the information based system has a more consistent performance than the conceptual distance system. It is very difficult to determine why there is this puzzling variance in performance across the test query set. At least one system produces reasonable results for all queries, effectively ruling out any argument concerning our query representations. What can be concluded from these results is that both systems have very different approaches to estimating semantic similarity and in future research it would make sense to combine both approaches in one system so as to take advantage of the good aspects of each approach and to compensate for weaknesses.

9.3 Full Test Evaluation

In this Section we present and discuss the results of a set of experiments to compare the performance of our semantic retrieval systems against that of a traditional pattern matching retrieval system. The pattern matching system used is the tf*IDF retrieval system described in Section 8.3. The evaluation mechanism is again precision/recall graphs, using the TREC relevance assessments and the WSJ text corpus. A fuller evaluation, (both with respect to the noise thresholding experiments and those presented here), would have been possible if we could have compared our systems performances against those of some of the systems used by the TREC participants. Unfortunately, however, detailed results in terms of complete rank lists were not made publicly available.

The remainder of this Section is organised as follows, in Section 9.1 we describe the experimental design and in Section 9.2 we present the results and discuss our findings.

9.3.1 Experimental Design

As in the experiments to determine the optimum noise thresholds, the amount of computing power required is a primary restraining factor in the design of these experiments. A direct comparison of both semantic approaches against the tf*IDF system over the entire WSJ collection was effectively impossible. Instead we opted to compare both systems on a set of 1,000 articles. The basic procedure involved the tf*IDF system retrieving the top 1000 documents in response to a query, and each of our systems was then used to rerank these documents with regard to their relevance to the query. As such, in place of a direct comparison we proposed using our approach to improve the results of the tf*IDF system.

In order to avoid any bias we used 12 queries, (this was as much as resources would permit). Once again there was no reason behind our decisions to choose

particular queries over others, beyond the fact that we decided not to reuse any of the queries from the thresholding experiments. Queries were both automatically and manually processed for the information based and conceptual distance systems in the same manner as described in Section 9.2.1. Queries for the tf*IDF system were made up of all terms from the original TREC query that were found to appear in less than 12% of the WSJ documents. The procedures involved in preprocessing documents for the information based and conceptual distance functions can be found in Chapter 7 and the equivalent operation for the tf*IDF system can be found in Section 8.3 of the previous Chapter.

9.3.2 Results

The performances of the conceptual distance, the information based, and the tf*IDF systems interpolated over the 12 test queries are presented in Figure 9-5. As can be seen, the traditional tf*IDF system out performs either of the semantic approaches, and, quite surprisingly, the information based semantic approach shows better performance than the conceptual distance system, (at least at high levels of recall). The average precision value for each system is shown in Table 9.6 :

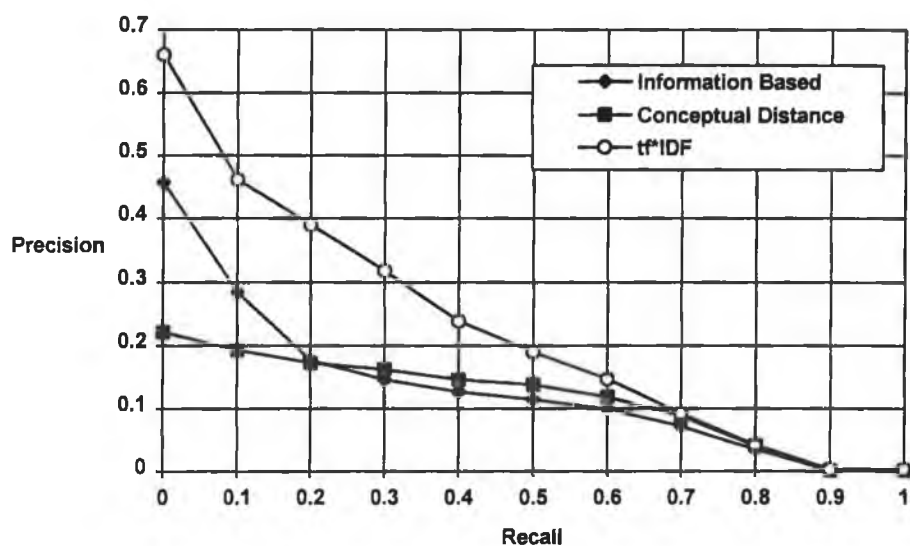
	Conceptual Dist.	Information Based	tf*IDF
Avg. Precision	0.1062	0.1151	0.2072

Table 9-7 Average Precision for all systems in the full test

A break down of the results on a query by query basis can be found in appendix J. From these results we see the tf*IDF system performs significantly better than either of the other approaches in queries 1, 2, 3, 7, 10, and 11. In query 5 the information based system shows better precision at low recall but quickly loses precision at higher level recall values, whereas the tf*IDF system maintains its level of precision for high recall levels. The same could be said in reverse with respect to the conceptual distance and tf*IDF systems in query 6. Only in query 8 did either semantic system

significantly outperform the $tf*IDF$ system. For this query the information based system performed very well at low recall, however, note again the severe decrease in precision at 0.2 recall.

(A)



(B)

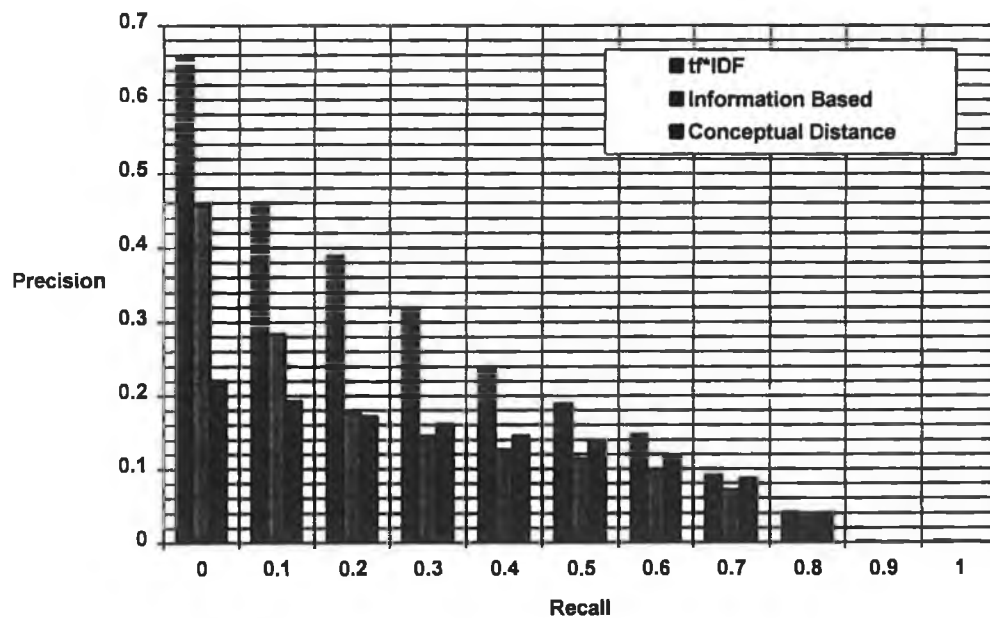


Figure 9-5 $tf*IDF$ Vs. Information Based Vs. Conceptual Distance

The performances of all three systems were very poor in queries 4 and 9, and although the semantic systems performed better in both queries, neither system managed to rerank the documents so as to put the relevant documents in the top 100 documents.

The obvious questions now are :

- why was there not an improvement over the $tf*IDF$ system ?
- why are the results for both semantic systems poorer than they were for the thresholding experiments ?
- why did the information based system perform better than the conceptual distance system when the opposite was the case in the thresholding experiments ?

We believe answers to these questions can be found in the combination of each of the following points :

(A) The TREC Method of assessing the relevant documents

TREC's method of determining relevant documents for a query could be said to be biased in favour of word-based pattern matching retrieval systems. Relevant documents are found by pooling the top 100 documents of a number of different retrieval systems. However, the backbone of most of these systems is a SMART-like pattern matching approach to IR. This would suggest that other pattern matching approaches to IR would perform well whereas different approaches as proposed by our information based and conceptual distance systems would not perform as well. We do not contend that the documents deemed relevant in TREC's relevance assessments are not relevant, but rather that they do not fully capture the complete set of relevant documents for a query. There is no reason to believe a user examining some of the documents ranked highly by our systems, but not by the $tf*IDF$ system, would not find these documents to be relevant to the query. In conclusion, a more complete, impartial computation of relevant documents called for. However, in fairness to

TREC's approach there is no obvious alternative. The idea of manually assessing 170,000 documents for relevance to 50 queries is quite clearly impractical⁵⁰.

(B) Our semantic based systems are not optimised

This point was made with regard to the results of the thresholding experiments, refer to Section 9.2.3, however, they are perhaps even more pertinent here. The $tf*IDF$ system is performance optimised for TREC queries on the WSJ collection by virtue of term weights. Features of the system such as the selective processing of posting lists and the cut-off threshold on the assignments of weights from the posting lists are specialities of the system we used and are further intended to improve performance. Figure 9-6 illustrates the effect of the cut-off threshold on performance over the 12 queries used in our full text evaluation.. In System A the cut-off threshold is set at 6,800, (this was found to be an optimal value and was used in our experiments), therefore after 6,800 articles are assigned a weight from the posting list, all new articles in subsequent postings are ignored. In system B the cut-off threshold was ignored. As can be seen the performance degrades in the absence of this threshold. The average precision for system B is 0.1517, this compares with 0.2007 for system A.

In contrast to the $tf*IDF$ system, there was no attempt to optimise either the information based or conceptual distance systems. In particular, there was no optimisation, or for that matter, evaluation of the sense disambiguator for either of the semantic retrieval systems. As was pointed out in [Sand94], and discussed in Section 7.2.1, any errors by the sense disambiguator often prove very costly in terms of retrieval performance. This point alone could account for the poor performances of our retrieval systems.

⁵⁰ Obviously the situation would be improved if we took part in TREC and included our ranked documents among those to be manually assessed.

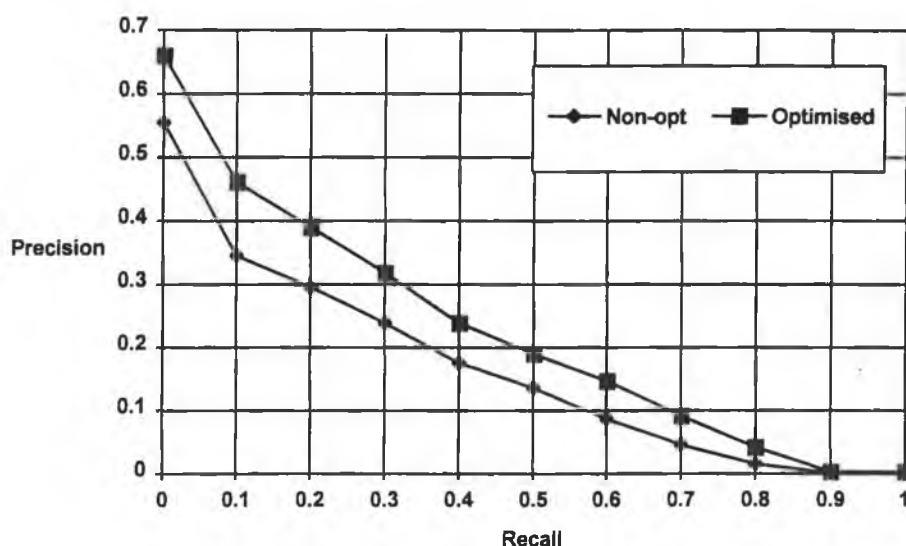


Figure 9-6 Optimised tf*IDF system Vs. Non-optimal tf*IDF system

(C) Existence of proper nouns

A characteristic of TREC queries is the existence of proper nouns such as company names, association and committee names, the names of people and the names of laws and programs. In general these terms are central to the query and make very good query terms. However, generally speaking they do not occur in WordNet and consequently cannot be used as query terms by our retrieval systems. Queries 1, 2, 3, 8, 10, and 11⁵¹ from our test query set include very many of these proper nouns (e.g. Commodity Futures Trading Commission CFTC, Inkatha, Chief Buthelezi, Ayatollah Khomeini, Shiite, etc.). It is not surprising that our systems performed particularly poorly in all of these queries.

All this would suggest that in future research an attempt should be made to extend our approach to facilitate the use of proper nouns, (of the kind just described), as query terms. Ginsberg, in his WorldViews system, [Gins93], combined a traditional

⁵¹ Note the overlap between these queries and those for which both semantic based systems performed very badly, 1, 2, 3, 7, 10, and 11.

pattern matching retrieval system with his knowledge based approach to information retrieval. Also, [Dee90], report on a process referred to as *latent semantic indexing* whereby term co-occurrence statistics are used to discover relationships between terms. We would propose that this co-occurrence information could be used to discover relationship between proper nouns. As such queries containing terms such as *Bill Gates*, *U2*, and *Spielberg* could be related to documents containing the terms *Microsoft*, *Bono*, and *ET* respectively. We would hypothesise that the best retrieval results could be obtained by a retrieval system adopting a semantic approach for non-proper noun terms, and a pattern matching approach extended to include latent semantic indexing for proper noun terms.

(D) Not a direct comparison

We believe the fact we did not have a direct comparison of retrieval systems but rather tried to use the information based and conceptual distance systems to improve the results of the $tf*IDF$ system, attributes to the poor performance of both semantic based systems. As pointed out in (A), documents deemed relevant by the $tf*IDF$ system might not be regarded as being relevant, (or at least not as relevant as other documents), by the semantic based retrieval systems. This relates back to the fact that documents may not have any index terms in common with a query yet may be relevant because they are indexed by terms related to query terms. Such documents are not included in the top 1000 returned by the $tf*IDF$ system.

One of the main strengths of our approach to information processing, the ability to relate semantically similar terms, is not fully afforded the opportunity to impact performance. By definition the documents making up the top 1000 documents of the $tf*IDF$ system does not include documents with terms related to the query, but rather documents with the actual query terms. We believe in an evaluation where all three systems were required to select the most relevant 1000 documents from the 153,000 documents in the WSJ, the two semantic retrieval systems would return many

more relevant documents. Unfortunately, due to current computational limitations and our present implementation, this was not possible.

(E) Form of TREC queries

The two problems with a non-semantic view of information, natural language ambiguity and the richness of natural language, are considerably reduced in the case of TREC queries. TREC queries are very detailed, including a concepts Section where the content of the query is expressed using many different terms and a definition Section where any ambiguous concepts are defined. This level of detail in queries is unusual in an information retrieval system and it is reasonable to assume that in a different test bed the performance of the tf*IDF system would be considerably poorer. We would expect the performances of the semantic systems to be unaffected by less detailed queries.

Our hypothesis regarding the richness of TREC queries is somewhat supported by the findings of other researchers. Query expansion, often used to increase the performance of traditional pattern matching retrieval systems has had little success in the TREC test bed. Voorhees, in [Voo94], reported very poor results in her attempts to increase the performance of SMART-like retrieval system by query expansion. Other factors contributed to her poor results, however, and we believe given the detail of TREC queries, expansion is unnecessary, and will more than likely disimprove results than yield an improvement.

(F) A representative test set

The question can be asked as to whether the 12 queries chosen are a representative query set. The performances of the semantic systems were considerably poorer over this query set than they were for the 12 queries used in the thresholding experiments. One of the reasons for this deterioration in performance could be related

to the fact we are dealing with lists of 1000 documents as opposed to 200 and the added 800 documents introduce noise which affects performance. We address this problem in (G). Another reason for the drop in performance could be the fact we simply chose a set of queries for which a semantic approach to processing did not suit. Certainly the points made in (C) would seem to support this theory. In an attempt to determine whether the query set *suites* the $tf*IDF$ retrieval system, we compared the performance of this system over the 12 test queries against its performance over the full 50 TREC-2 queries, as figure 9.7 illustrates. As can be seen, the performance is marginally better for the test query set, however, overall the performance of the $tf*IDF$ system over the test query set would have to be said to be representative of the system's performance in general. The average precision for the $tf*IDF$ system over 50 queries is 0.1916 as compared with 0.2007 for the 12 queries.

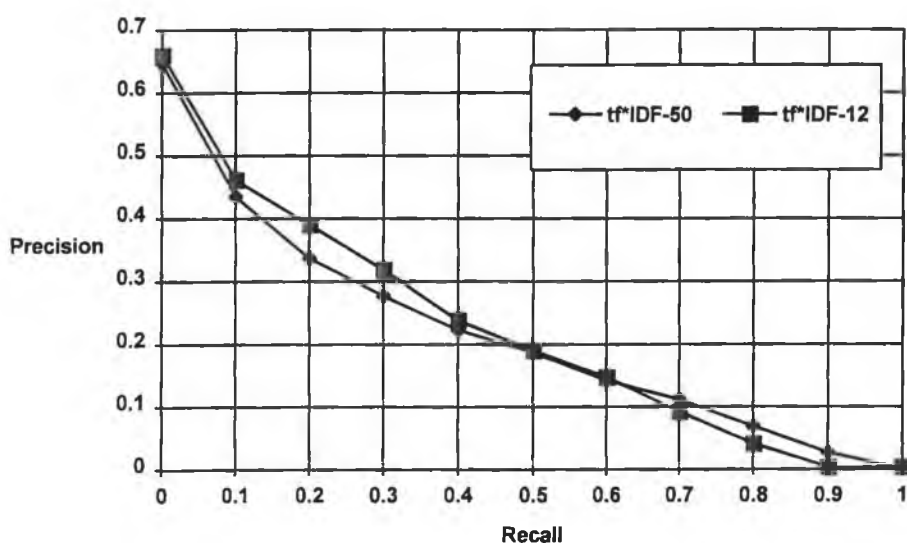


Figure 9-7 Performance of the $tf*IDF$ system over 12 and 50 queries

(G) The effect of noise in the larger document set

In (H) we made the point that the extra 800 documents in the document sets of the full test experiments could introduce noise which would account for the

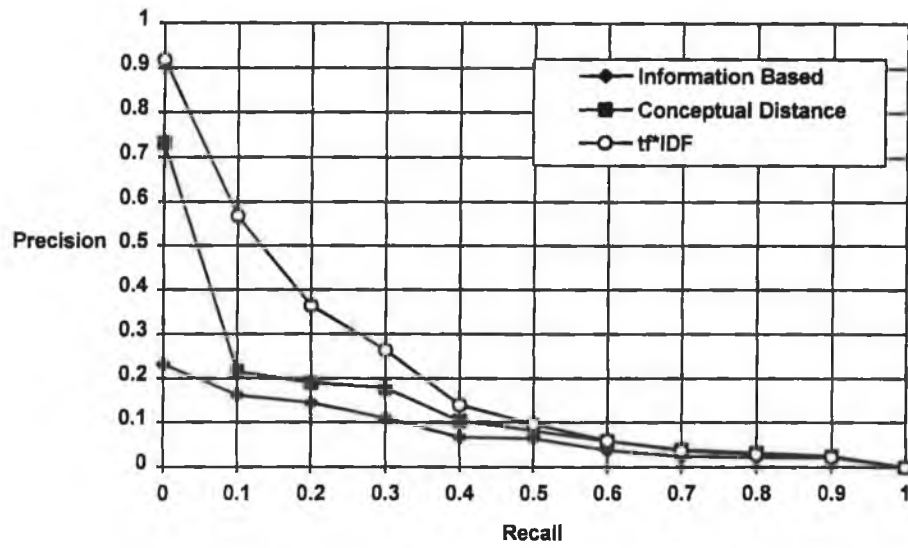
deterioration in retrieval performance. One could also hypothesise that the conceptual distance system is less successful in dealing with noise than the information based system. This would then explain why the conceptual distance system went from performing slightly better than the information based system to a situation where it performed significantly poorer. In an attempt to resolve these two issues we ran four more tests. For these runs we used four of the best performing queries from the thresholding experiments, queries 4, 5, 10 and 11. From appendix J it can be seen that the conceptual distance system out performed the information based system for each of these queries. The organisation of these runs was exactly the same as for the 12 full test queries, using 1000 documents returned by the tf*IDF system as a test bed and not altering the queries from the thresholding experiments. The results are presented in Table 9.7 and Figure 9-8, (a break down of results for each query can be found in appendix J).

Once again the tf*IDF system performed better than both semantic systems, out performing both systems in all queries except query 10, where the conceptual distance system performed marginally better. However, of particular interest in these runs is the fact that the conceptual distance system performed significantly better than the information based system. This consistent performance of the conceptual distance system rules out any argument of it suffering from the effects of greater noise in a larger test bed. It also indicates that the set of 12 queries used in the full test evaluation were not representative of the conceptual distance system's best performance, and with a different set, the performance could be better.

	Conceptual Dist.	Information Based	tf*IDF
Avg. Precision	0.1078	0.0753	0.1921

Table 9-8 Average Precision for all systems on 4 best thresholding queries

(A)



(B)

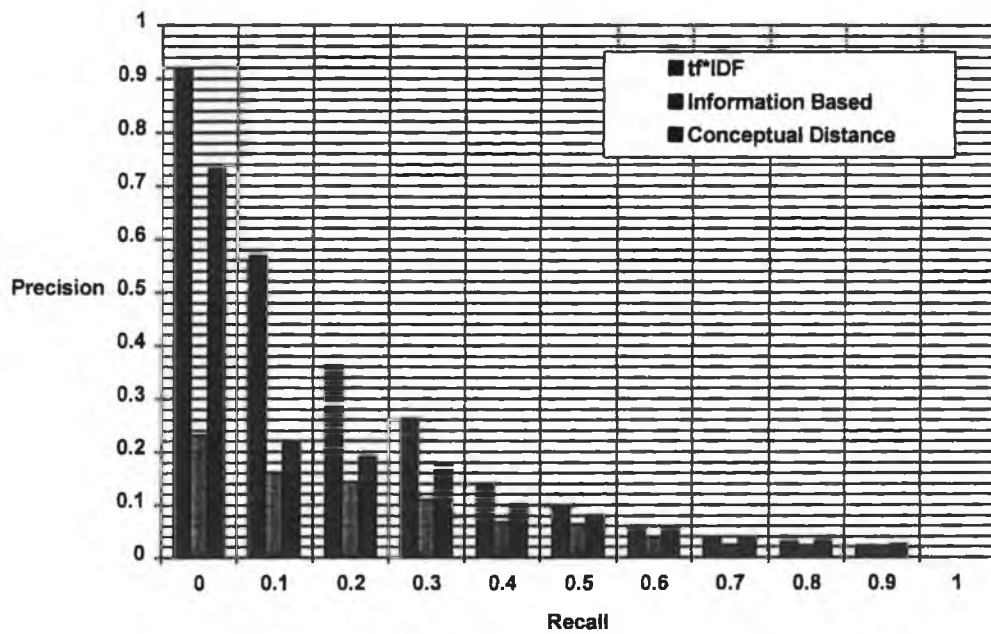


Figure 9-8 Performance of all three systems on 4 best thresholding queries

9.3.3 Conclusions on evaluation

We will present overall conclusions to our research in the next Chapter, however, from the experiments described here and in Section 9.2 we can broadly conclude that our approach to information retrieval shows promise, but the true potential of our approach has yet to be determined. Further research is required to determine an optimum configuration for both the information based and conceptual distance systems. In particular, future work is required to determine the degree of accuracy of our sense disambiguator. At present, when presented with the overall performance of either semantic system, it is difficult to distinguish between the performance of the retrieval engine and the sense disambiguator.

What does seem apparent from our experiments is the fact that the information based and conceptual distance approaches to semantic similarity have different strengths and weaknesses. This can be seen in the varying performances of both retrieval systems for the same queries. A combined retrieval system, using the strengths of both approaches would appear to be the obvious course of action, however, more research is required to isolate these strengths. Another improvement highlighted by our experiments is a facility to make use of proper noun terms as both index and query terms. In regard to this improvement it was proposed that latent semantic indexing could be used to relate proper noun index and query terms, thus further improving performance.

A final lesson learned from our experiments is the fact that the TREC environment does not suit our approach to information retrieval. TREC queries are uncharacteristically detailed, effectively curbing the problems posed by natural language ambiguity and complexity, the two weaknesses of traditional IR systems addressed by our approach. It is also known that the set of relevant documents, as proposed by TREC, is a subset of the true set of relevant documents. The pooling mechanism of determining relevant documents is flawed by using the results of predominantly pattern matching retrieval systems, (though, as we point out, there is little that can be done about this). Finally our approach to evaluating performance,

reranking 1000 documents retrieved by the $tf*IDF$ retrieval system, could well have been better carried out by a direct comparison approach over all documents. However, in defence of our approach to the evaluation we can put forward arguments regarding the restrictions imposed by processing power required and the absence of other suitable test beds and evaluation mechanisms, (refer to Chapter 6).

9.4 Summary

In this Chapter we presented the results of an evaluation of our approaches to information retrieval. Prior to the evaluation it was necessary to determine values for the absolute noise threshold variable. This was accomplished using a set of 12 TREC queries and sets of 200 documents from the WSJ. The evaluation itself involved the use of the semantic retrieval systems in a process to rerank 1000 documents returned by a tf/DF retrieval system in response to a TREC query. 12 such runs were carried out. The results of this evaluation were discussed in Section 9.3.2 and suggestions for possible improvements as well as alternative approaches to future evaluations were proposed.

In the following Chapter the results of our approach to information retrieval, and information processing in general, are summarised. The Chapter will also propose directions for future work in all the areas of research addressed in this thesis.

Chapter 10 - Conclusions and Future Directions

10. Introduction

The research reported in this thesis has centred around the development of a semantic based approach to information processing to replace the traditional word-based pattern matching approach. Our proposed semantic information processing system was comprised of a WordNet derived, domain independent KB and a concept level semantic similarity estimator. The KB was used as a controlled vocabulary which effectively addressed many of the problems posed by ambiguities in natural language. Similarly both proposals for the semantic similarity estimator tackle issues regarding the richness of natural language and in particular the multitude of ways of expressing the same concept.

A semantic based document retrieval system was developed as a means of evaluating our approach. However, many other applications were discussed with particular attention directed towards the application of our approach to locating and relating information in a large scale FDBS, (refer to Sections 1.1 and 6.1). The document retrieval evaluation application entailed the development of an automatic sense disambiguator, (Section 7.2). Our evaluation mechanism was to use the Wall Street Journal text corpus and a set of TREC queries along with their relevance assessments. A traditional pattern matching $tf*IDF$ system was used as a baseline system in our evaluation experiments. The basic procedure involved obtaining KB representations of both the documents and queries and using the semantic similarity estimators as the comparison mechanism in the procedure to determine the degree of relevance of a document for a query.

The results of our experiments showed that neither the information based or the conceptual distance retrieval systems were as good as the conventional $tf*IDF$ system.

However, we believe the arguments put forward in Section 9.3.3 explain that these results should not be seen as wholly negative but rather as offering promise for the future. Many of our queries perform very well with our strategy and notwithstanding the reasons given in Section 9.3.3, we believe our results are certainly worth pursuing. We would broadly conclude from our experimental evaluation that :

- WordNet can be used as the basis for a domain independent controlled vocabulary in an information processing task.
- Our approach to the task of automatically weighting relational links in large scale concept graphs would appear to produce good results in the estimation of conceptual distance
- Including collocations in the calculation of information content values improves the information based estimator of semantic similarity.
- Both the information based and conceptual distance similarity estimators are reasonable computational estimations of the semantic similarity between concepts
- Our approach to automatic word sense disambiguation is at the very least a good starting point for future research in this area.

Finally, we believe the application of semantic information processing system in the less well defined environments of Non-Self Describing (NSD) applications would produce better results. Certainly we would argue that our strategy compares very favourably with any existing system in what we describe as NSD applications.

10.1 Future Directions

In this Section we will briefly outline some of the research directions we believe should be further pursued as a result of the research reported in this thesis.

This list is in no sense exhaustive and is merely intended to highlight the larger areas that could be investigated.

(A) Knowledge base

Further work is required into our KB construction process. In particular there is a definite need to have the KB resident in main memory. This could be accomplished by some intelligent paging mechanism or perhaps by redefining the set component HCGs so as to have a larger number of smaller HCGs. Other future developments could include extensions to the KB to include the adjective and verb Sections of WordNet. However, this is a particularly ambitious task and future extensions to the KB should probably be restricted to additions in response to extensions to WordNet made by the WordNet developers at Princeton.

(B) Weighting Mechanism

Future research on our automatic weighting mechanism could obviously include a more complete evaluation. This would allow a more complete understanding of the contributions of the component parts, (*depth scaling*, *local density* and *strength of connotation*), and perhaps the effects of varying their constituent parameters.

(C) Similarity Estimators

It is apparent from the evaluation experiments presented in Chapter 9, that the information based and conceptual distance similarity estimators are sufficiently different in their approach to warrant a future investigation into proposals for combining them. Again a fuller evaluation of each approach would help to resolve such questions as; What is the contribution of the non-hierarchical links to the effectiveness of conceptual distance similarity estimator? Is information based

similarity estimator developed using the WSJ text corpus sufficiently domain independent for use in other applications?

(D) Text Preprocessor

In Section 7.1 we investigated approaches to removing non-content bearing terms from text. Further research could be carried out in this area. Interesting avenues could include the following; Can information content values be used to eliminate non-content bearing terms ? Should proper nouns that do not appear in the KB be stripped out?

(E) Sense Disambiguator

The sense disambiguator developed in this research was a fairly substantial undertaking and represented a new approach to word sense disambiguation, however, there was no formal evaluation. This made it difficult to determine whether our poor results were as a result of the poor performance of the sense disambiguator or the retrieval engine, or perhaps a combination of both for different queries. Up until recently it has been very difficult to evaluate sense disambiguators, however, the manual tagging of an extract of the Brown Corpus with WordNet synsets, [Mill94], should provide an excellent test bed for an evaluation of our approach to sense disambiguation. Proposed improvements to our disambiguation strategies, suggested in Section 7.3, could subsequently be implemented.

(F) Document Retrieval Application

In Section 9.3.3 we proposed possible improvements to our semantic-based document retrieval systems. One of these improvements, the extension to facilitate a mechanism for latent semantic indexing, we believe will produce very much improved

results. As well as changes to the document retrieval system we would like to see semantic-based systems developed for other SD applications. In particular a semantic based information filtering system would be a worthwhile future research project.

(G) NSD Application

Finally, we believe future research should be carried out into applying our semantic information processing strategies to an NSD application. It may be possibly to use Bright's FDBS simulator, [Brig94], in an FDBS application, or alternatively, develop a multimedia database to use as a test bed in a multimedia NSD application.

References

- [Bars92] : L. W. Barsalou, "Cognitive Psychology : An overview for Cognitive Scientists", Cognitive Science Series, Tutorial Essays, 1992.
- [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 Teng, "Design and Implementation of the WordNet Lexical Database and Searching Software", Working Paper, Princeton University, 1993.
- [Boo91] : G. Booch, "Object Oriented Design with Applications", the Benjamin/Cummings Publishing Company, 1991.
- [Bota92] : A. Botafogo, E. Rivlin, and B. Shneiderman, "Structural Analysis of Hypertexts: Identifying Hierarchies and useful Metrics", ACM Transactions on Information Systems, Vol. 10, No. 2, April 1992, 142 - 180.
- [Boug92a] : A. Bouguettaya and R. King, "Large Multidatabases : Issues and Directions", IFIP, DS-5 Semantics of Interoperable Database Systems, Volume 1, Australia, Nov 16-20, 1992, 53-64.
- [Boug92b] : A. Bouguettaya, "A Dynamic Framework for Interoperability in Large Multidatabases", PhD. Thesis, Colorado University, May 1992.
- [Bril93] : E. Brill, "A simple rule-based part of speech tagger", Proceedings Speech and Natural Language Workshop, San Mateo, CA, February 1992, Morgan Kaufman, 112-116.

[Brig92] : M. W. Bright, A. R. Hurson, and S. H. Pakzad, "A Taxonomy and Current Issues in Multidatabase Systems", IEEE Computer, March 1992, 50-59.

[Brig94] : M. W. Bright, A. R. Hurson, and S. Pakzad, "Automated Resolution of Semantic Heterogeneity in Multidatabases", ACM Transaction on Database Systems, Vol. 19, No. 2, June 1994, 212 - 253.

[Buck85] : C. Buckley : "Implementation of the SMART Information Retrieval System", Tecnical Report 86-686, Computer Science Department, Cornell University, Ithaca, New York, May 1985.

[Cat94] : W. L. Cats-Basil and T. Jelassi, "The French Videotex System Minitel : A successful implementation of a national information technologies infrastructure", MIS Quarterly, Vol. 18, No. 1, March 1994, 1-20.

[Chen92] : H. Chen and K. Lynch, "Automatic Construction of Networks of Concepts Characterizing Document Databases", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 22, No. 5, September/October 1992, 885 - 902.

[Crof91] : W. B. Croft, H. R. Turtle, and D. D. Lewis, "The use of Phrases and Structured Queries in Information Retrieval", Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, Oct 13 - Oct 16, 1991, 32-45.

[Crof93] : W. B. Croft, "Knowledge Based and Statistical Approaches to Text Retrieval", IEEE Expert, April 1993, 8-12.

[Dee90] : S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis", Journal of the American Society of Information Science, Vol. 41, No. 6, 1990, 391 -407.

[Dunn93] : C. Dunne, "Exploiting a User Model to Dynamically Generate Guided Tours in Hypertext", MSc. thesis, School of Computer Applications, Dublin City University, 1993.

[Evan91] : D. A. Evans, "Automatic Indexing Using Selective NLP and First Order Thesauri", Proceedings of RIAO '91, Barcelona, Spain, 1991, 624 - 643.

[Felb90a] : Christiane Felbaum, Derek Gross, and Katherine Miller, "Adjectives in WordNet", International Journal of Lexicography, Vol. 3, No. 4, 1990.

[Felb90b] : Christiane Felbaum, "English Verbs as a Semantic Net", International Journal of Lexicography, Vol. 3, No. 4, 1990.

[Fox90] : C. Fox, "A Stop List for General Text", SIGIR Forum. Vol. 24, No 1-2, Fall 89 / Winter 90, 1990, 19 - 35.

[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.

[Guin93] : C. A. Guinan, "An Investigation into Automatic Route Selection in a Hypertext Used for Learning", MSc. thesis, School of Computer Applications, Dublin City University, 1992.

[Harm93] : D. Harman, "An Overview of the First Text Retrieval Conference (TREC-1)", TREC 2, 1993, 1-20.

[Hear93] : M. A. Hearst and C. Plaunt, "Subtopic Structuring for Full-Length Document Access", Proceedings of the Sixteenth Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval, Pittsburgh, June 27 - July 1, 1993, 59-68.

[Heim85] : D. Heimbigner and D. McLeod, "A Federated Architecture for Information Management", ACM Transactions on Office Information Systems 3, 3 (July) 1985, 253-278.

[Henl69] : N. M. Henley, "A Psychological Study of the Semantics of Animal Terms", Journal of Verbal Learning and Verbal Behavior, 8, 1969, 176 - 184.

[John88] : P. N. Johnson-Laird, "The Computer and the Mind : An Introduction to Cognitive Science", Harvard University Press, 1988.

[Karl89] : F. Karlsson, "Designing a Parser for Unrestricted Text", Unpublished Manuscript, 1989.

[Kim90] : Young Whan Kim and Jinh H. Kim, "A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph", Journal of Documentation, Vol. 46, No. 2, June 1990, 113-137.

[Krov92] : R. Krovetz and W. B. Croft, "Lexical Ambiguity in Information Retrieval", ACM Transactions on Information Systems, Vol. 10, No. 2, April 1992, 115-141.

[Lee93] : J. H. Lee, M. H. Kim, and Y. J. Lee, "Information Retrieval Based on Conceptual Distance in IS-A Hierarchies", Journal of Documentation, Vol. 49, No. 2, June, 1993, 113 -136.

[Lehm75] : E. L. Lehman, "Nonparametrics : Statistical Methods Based on Ranks", Holden-Day series in Probability and Statistics, McGraw-Hill Publishing Company, 1975.

- [Lesk86] : M. Lesk, "Automatic Sense Disambiguation : How to tell a pine cone from an ice cream cone", Proceedings of the SIGDOC Conference, 1986; 24-26.
- [McCa80] : D. B. McCarn, "MEDLINE : An Introduction to On-line Searching", Journal of the American Society for Information Science, 31(3), 1980, 181 - 192.
- [Meye70] : D. E. Meyer, "On the Representation and Recall of Stored Semantic Information", Cognitive Psychology, 1970, 1, 242-300.
- [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.
- [Mill91] : G. A. Miller and G. Charles, "Contextual Correlates of Semantic Similarity", Language and Cognitive Processes, 6(1), 1991, 1-28.
- [Mill93] : George A. Miller, "Nouns in WordNet : A Lexical Inheritance System", Internal Report., Cognitive Science Laboratory, Princeton University, Princeton, NJ, August 1993.
- [Mill94] : G. A. Miller, C. Leacock, R. Tangi, and R. T. Bunker, "A Semantic Concordance", Internal Report, Cognitive Science Laboratory, Princeton University, Princeton, NJ, 1994.
- [Mona87] : I. Monarch and J. Carbonell, " CoalSORT : A Knowledge-Based Interface", IEEE Expert, Spring 1987, 39 - 53.
- [Niel89] : J. Nielsen, "Hypertext and Hypermedia", Academic Press Inc., 1989.

- [ODon94] : R. O'Donnell and A. F. Smeaton, "Re-Implementation of SIMPR TSAs", BCS-IRSG colloquium, Glasgow, March 22-23, 1994.
- [Paic91] : C. Paice, "A Thesaural Model of Information Retrieval", *Information Processing and Management*, Vol. 27, No. 5, 1991.
- [Pitk94] : J. E. Pitkow and M. Recker, "Results from the first World Wide Web user survey", *Proceedings of the First International World-Wide Web Conference*, Geneva, May 25-27, 1994, 283-294.
- [Poli87] : S. Pollitt : "CANSEARCH : An Expert System Approach to Document Retrieval", *Information Processing and Management*, Vol. 23, No. 2, 1987, 119 - 138.
- [Port80] : M. F. Porter, "An Algorithm for Suffix Stripping", *PROGRAM*, Vol. 14, No. 3, 1980, 130 -137.
- [Quin68] : M. R. Quillian, "Semantic Memory", in *Semantic Information Processing*, M. Minsky, (Ed.). Cambridge, MA: MIT Press, 1968.
- [Rada88] : R. Rada and H. Mili, "Merging Thesauri: Principles and Evaluation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 2, March 1988, 204 -220.
- [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.

[Rich94a] : R. Richardson, A. F. Smeaton, and J. Murphy, "Using WordNet for Conceptual Distance Measurement", BCS-IRSG colloquium, Glasgow, March 22-23, 1994.

[Rich94b] : R. Richardson, A. F. Smeaton, and J. Murphy, "Using WordNet as a Knowledge Base for Measuring Conceptual Similarity between Words", Proceedings of the Seventh Annual Conference on Artificial Intelligence and Cognitive Science, (AICS), 1994, 179 -192.

[Rips73] : L. J. Rips, E. J. Shoben, and E. E. Smith, "Semantic Distance and the Verification of Semantic Relations", Journal of Verbal Learning and Verbal Behavior, 8, 1973, 1 - 20.

[Rube65] : H. Rubenstein and J. B. Goodenough, "Contextual Correlates of Synonymy", Computational Linguistics, 8, 1965, 627-633.

[Salt83] : G. Salton and M. McGill, "Introduction to Modern Information Retrieval", Computer Science Series, McGraw-Hill, Inc., 1983.

[Salt89] : G. Salton, "Automated Text Processing - The Transformation, Analysis, and Retrieval of Information by Computer", Addison-Wesley Series in Computer Science, 1989.

[Salt90] : G. Salton, "On the Application of Syntactic Methodologies in Automatic Text Analysis", Information Processing and Management, 26(1), 1990, 73 - 92.

[Salt93] : G. Salton, J. Allan, and C. Buckley, "Approaches to Passage Retrieval in Full Text Information Systems", Proceedings of the Sixteenth Annual International

ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, June 27 - July 1, 1993, 49-57.

[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, July 3 - July 6, 1994, 49-57.

[Schw90] : C. Schwarz, "Content Based Text Handling", Information Processing and Management, Vol. 26, No. 2, 219-226, 1990.

[Sedg88] : R. Sedgewick, "Algorithms", Second Edition, Addison-Wesley, 1988.

[Seki93] : J. Sekine, M. Kawashimo, and M. Nakagawa, "An Architecture of the Data Standardization Tool that uses a Word Dictionary", Paper submitted to VLDB '93, 1993.

[Shep62a] : R. N. Shepard, "The Analysis of Proximities : Multidimensional Scaling with an unknown Distance Function", Part I, Psychometrika, 27, 1962, 125 - 140.

[Shep62b] : R. N. Shepard, "The Analysis of Proximities : Multidimensional Scaling with an unknown Distance Function", Part I, Psychometrika, 27, 1962, 219 - 246.

[Sher92] : P. Sheridan and A. F. Smeaton, "The Application of Morpho-syntactic Language Processing to effective Phrase Matching", Information Processing and Management, Vol. 28, No. 3, 1992.

[Shet90] : A. P. Sheth and J. A. Larson, "Federated Database Systems for managing Distributed, Heterogeneous, and Autonomous Databases", ACM Computing Surveys, Vol. 22, No. 3, Sept 1990, 183-236.

[Shov85] : P. Shoval, "Principles, Procedures and Rules in an Expert System Approach for Information Retrieval", *Information Processing and Management*, Vol. 21, No. 6, 1985, 475-487.

[Silb91] : A. Silberschatz, M. J. Stonebraker, J. Ullman, "Database Systems : Achievements and Opportunities", *Communications of the ACM*, Vol. 34, No. 10, October 1991, 110-120.

[Smar92] : G. Smart, *The SIMPR Project: The Results*, CEC Publication:SIMPR-CRI-1992-41.10e, 1992.

[Smea90] : A. F. Smeaton, "Hypertext : The Past the Present and our Future", School of Computer Applications, Dublin City University, Working Paper : CA0290.

[Smea91] : A. F. Smeaton, "Prospects for Intelligent, Language Based Information Retrieval", *Online Review*, Vol. 15, No. 6, 1991, 373-382.

[Smea92a] : A. F. Smeaton, "Progress in the Application of Natural Language Processing to Information Retrieval Tasks", *The Computer Journal*, Vol. 35, No. 3, 1992, 268 -278.

[Smea92b] : A. F. Smeaton, "An Evaluation of Retrieval Performance Using Simple Statistics and SIMPR Linguistic Processing an a Standard Collection of Texts", CEC Publication: SIMPR-DCU-1992-50.2i, 1992.

[Smea94] : A. F. Smeaton, "Linguistic Approaches to Text Management: An Appraisal of Progress", to appear in the *Journal of Document and Text Management*.

[Spar75] : K. Sparck Jones and C. J. van Rijsbergen, "Report on the need for and Provision of an "Ideal" Information Retrieval Collection", *British Library Research and Development Report 5266*, Computer Laboratory, University of Cambridge, 1975.

[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.

[Sutcl91] : R. Sutcliffe, "Distributed Representations on a text based Information Retrieval System : A new way of using the Vector Space Model", Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, Oct 13 - Oct 16, 1991, 123-132.

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

[Tver78], : A. Tversky and I. Gati, "Studies of Similarity", in *On the Nature and Principle of Formation of Categories*, E. Rosch and B. Lloyd, (Eds.). Hillsdale, N. J. : Erlbaum, 1978.

[VanR79] : C. J. Van Rijsbergen, "Information Retrieval", Second Edition, Butterworth and Co. Publishers, 1979.

[Voor93] : E. M. Voorhees, "Using WordNet to Disambiguate Word Senses for Text Retrieval", Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, June 27 - July 1, 1993, 171-180.

[Voor94] : 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, July 3 - July 6, 1994, 61 - 69.

[Vout92] : A. Voutilainen, J. Heikkilä, A. Anttila, "Constraint Grammar of English", University of Helsinki Publication No. 21, ISBN 951-45-6276-3, 1992.

[Will94] : P. Willett and P. Ingwersen, "An Introduction to Information Retrieval", Tutorial Notes, Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, July 3 - July 6, 1994.

[Wins87] : M. E. Winston, R. Chaffin, and D. J. Hermann, "A Taxonomy of Part-Whole Relations", Cognitive Science, 11, 417-444.

[Yaro93] : D. Yarowsky, "One Sense per Collocation", Proceedings of ARPA Human Technology Workshop, 1993.

Appendix A - HCG Construction

This appendix contains the results of a set of experiments to evaluate the coverage and degree of overlap for two different sets of HCG roots. Refer to section 4.1 for further discussion. The first set of roots had 22 root concepts and given the difficulty in presenting a 22 x 22 matrix, it was decided to split it up and present three matrices of size 22 x 7, 22 x 7, and 22 x 8 respectively. Each cell entry contains the percentage of unique concepts in the row HCG that are found in the column HCG. The asterisks are used to highlight high degrees of overlap.

	Act	Animal	Artifact	Attribute	Cognition.	Comm.	Event
Act	-	000.02	003.22	004.17	003.99	005.34	006.57
Animal	000.01	-	000.01	000.00	000.01	000.00	000.00
Artifact	001.50	000.01	-	000.34	000.24	000.66	000.34
Attribute	006.32	000.00	000.98	-	004.35	002.47	001.37
Cognition	007.53	000.04	000.87	005.44	-	005.02	001.52
Communication	005.94	000.00	001.75	001.61	003.14	-	000.78
Event	025.30**	000.00	002.80	003.90	003.31	002.89	-
Feeling	012.00	000.00	000.35	023.88**	005.65	003.53	000.71
Food	000.54	000.32	000.29	000.29	000.11	000.04	000.11
Group	000.04	000.12	000.01	000.07	000.11	000.00	000.05
Location	000.37	000.00	000.57	000.04	000.04	000.29	000.16
Object	000.24	000.06	001.59	000.37	000.00	000.49	000.43
Person	000.28	000.07	000.72	000.25	000.21	000.39	000.26
Phenomena	004.17	000.00	000.58	001.55	000.87	000.19	002.52
Plant	000.00	000.00	000.08	000.17	000.00	000.00	000.00
Possession	006.84	000.12	002.78	002.44	000.93	003.83	000.70
Process	005.70	000.00	000.66	001.54	000.88	000.00	003.07
Quantity	001.40	000.00	001.70	001.32	000.43	000.64	000.51
Relation	007.78	000.00	001.99	003.61	003.54	076.81**	001.59
Shape	006.50	000.00	006.67	001.88	001.71	001.54	003.25
State	005.75	000.00	000.48	006.75	001.67	001.11	001.85
Substance	000.49	000.19	000.62	000.34	000.06	000.06	000.17

	Feeling	Food	Group	Location	Object	Person	Phenomen
Act	001.89	000.34	000.05	000.23	000.07	000.34	000.96
Animal	000.00	000.13	000.13	000.00	000.01	000.06	000.00
Artifact	000030	000.08	000.01	000.13	000.26	000.36	000.05
Attribute	006.35	000.24	000.15	000.03	000.18	000.45	000.54
Cognition	001.71	000.11	000.30	000.04	000.00	000.49	000.30
Comm.	000.72	000.02	000.00	000.18	000.25	000.43	000.07
Event	000.42	000.34	000.34	000.34	000.51	001.27	002.21
Feeling	-	000.47	000.00	000.00	000.35	000.12	001.06
Food	000.11	-	000.18	000.00	000.29	000.32	000.00
Group	000.00	000.07	-	000.03	000.01	000.44	000.00
Location	000.00	000.00	000.08	-	000.21	000.08	000.04
Object	000.12	000.49	000.06	000.31	-	000.55	000.00
Person	000.02	000.16	000.60	000.04	000.16	-	000.00
Phenomena	000.68	000.00	000.00	000.10	000.00	000.00	-
Plant	000.00	000.29	006.42	000.00	000.00	000.13	000.13
Possession	000.12	000.46	000.12	000.35	000.00	000.35	000.23
Process	000.44	000.00	000.00	000.00	000.00	000.00	100.00**
Quantity	000.09	000.21	00.017	000.04	000.38	000.51	000.00
Relation	001.30	000.03	000.08	000.17	000.22	000.68	000.20
Shape	000.34	000.17	000.00	000.00	014.53	001.03	000.51
State	003.45	000.22	000.00	000.11	000.04	000.07	001.26
Substance	000.06	059.46**	000.28	000.04	000.36	000.23	000.06

	Plant	Poss.	Process	Qty.	Relat.	Shape	State	Subst.
Act	000.00	001.48	000.62	000.68	008.69	000.80	003.56	000.57
Animal	000.00	000.01	000.00	000.00	000.00	000.00	000.00	000.13
Artifact	000.02	000.21	000.03	000.40	000.93	000.45	000.12	000.27
Attribute	000.12	000.66	000.27	000.98	006.67	000.33	005.96	000.48
Cognition	000.00	000.42	000.15	000.34	006.77	000.53	001.90	000.11
Comm.	000.00	000.94	000.00	000.43	099.91*	000.20	000.74	000.04
Event	000.00	000.51	001.10	001.02	006.79	001.61	004.07	000.76
Feeling	000.00	000.12	000.24	000.24	008.12	000.24	010.71	000.47
Food	000.29	000.14	000.00	000.14	000.07	000.04	000.18	099.89*
Group	002.04	000.01	000.00	000.05	000.07	000.00	000.00	000.17
Location	000.00	000.12	000.00	000.04	000.37	000.21	000.12	000.08
Object	000.00	000.00	000.00	000.49	000.61	005.14	000.06	000.98
Person	000.07	000.07	000.00	000.19	000.78	000.12	000.04	000.19
Phenomena	000.29	000.19	045.05*	000.00	001.07	000.29	003.20	000.29
Plant	-	000.00	000.13	000.04	000.00	000.00	000.08	000.80
Possession	000.00	-	000.12	001.28	005.80	000.23	001.16	000.58
Process	000.66	000.22	-	000.00	001.54	000.22	003.95	000.66
Quantity	000.04	000.47	000.00	-	000.94	000.51	001.91	001.11
Relation	000.00	001.08	000.10	000.46	-	000.25	001.35	000.14
Shape	000.00	000.34	000.17	001.54	002.56	-	001.17	000.17
State	000.07	000.30	000.67	001.41	002.60	000.33	-	012.16
Substance	000.43	000.11	000.06	000.36	000.19	000.02	007.04	-

There were 11 HCGs in the second set of experiments. Again due to the difficulty of presenting an 11 x 11 matrix, it was decided to split it up and present two matrices of size 11 x 6 and 11 x 5 respectively. Each cell again contains the percentage of unique concepts in the row HCG that are found in the column HCG. The bracketed numbers in the first column of the first table indicate the number of concepts in each HCG.

		Abstract.	Act	Psyc_feat.	Entity	Event	Group
Abstraction	(11781)	-	005.15	003.80	002.05	000.81	000.10
Act	(7623)	007.98	-	003.74	002.64	004.34	000.04
Psyc_feature	(3837)	010.71	006.67	-	000.78	001.02	000.23
Entity	(43950)	000.54	000.47	000.08	-	000.11	000.45
Event	(1311)	006.33	021.51	002.9	003.36	-	000.15
Group	(11318)	000.09	000.03	000.08	001.72	000.02	-
Location	(2664)	000.26	000.41	000.04	000.79	000.11	000.08
Phenomenon	(1190)	001.68	002.77	001.18	001.26	002.77	000.00
Possession	(964)	005.39	005.39	000.73	002.80	000.52	000.10
Shape	(688)	003.20	005.81	000.87	027.18	002.62	000.00
State	(2506)	008.18	005.67	005.19	000.84	001.76	000.00

	Location	Phenome.	Poss.	Shape	State
Abstraction	000.07	000.19	000.56	000.23	002.01
Act	000.17	000.49	000.89	000.62	002.24
Psyc_feature	000.03	000.42	000.21	000.16	003.57
Entity	000.05	000.04	000.07	000.46	000.05
Event	000.23	002.52	000.38	001.30	003.28
Group	000.02	000.00	000.01	000.00	000.00
Location	-	000.04	000.11	000.19	000.04
Phenomenon	000.08	-	000.08	000.34	001.85
Possession	000.31	000.10	-	000.10	000.62
Shape	000.58	000.58	000.15	-	001.45
State	000.04	000.84	000.20	000.36	-

Appendix B - Calculating Information Content Values

This appendix presents details on the results of calculating information content values using the Wall Street Journal text corpus. There was no handling of collocations in this implementation.

HCG Name	A	B	C	D	E	F	G	H
Abstraction	6637	61%	45%	90%	5.57	6.23	1.62	1.10
Act	4301	65%	42%	95%	4.62	5.28	1.54	1.10
Entity	24424	41%	70%	91%	6.16	7.52	1.64	1.14
Event	727	69%	53%	94%	4.16	5.07	1.40	1.03
Group	915	15%	57%	92%	4.89	5.13	1.40	1.01
Location	1034	64%	62%	96%	5.21	5.06	1.19	1.04
Phenomenon	725	47%	57%	87%	4.56	4.68	1.22	1.07
Possession	613	66%	74%	96%	4.34	4.79	1.16	1.08
Psych_feature	2184	64%	47%	94%	5.33	6.50	1.40	1.05
Shape	402	58%	54%	89%	3.47	4.27	1.30	1.04
State	1413	60%	48%	93%	3.78	5.56	1.20	1.05

A - Total Synsets

B - Total % of Synsets that Get an Information Content Value

C - Percentage of Concepts from Synsets that Don't Receive a Value that are Collocations

D - Percentage of Synsets Recieving a Value that are Directly Found

E - Average Depth of Directly Found Synsets

F - Average Depth of Unfound Synsets

G - Average Polysemy of Directly Found Synsets

H - Average Polysemy of Unfound Synsets

Appendix C - Senses Used in Psychological Evaluation

In this appendix we present the sense of words used by the conceptual distance and information based similarity estimators in the psychological evaluation. Table 5.1 from section 5.3 lists the 28 noun pairs used in the evaluation process. For each ambiguous word in a noun pair the different senses are presented and the one used in the evaluation is highlighted in bold. The amount of ambiguity in this relatively small set of nouns gives a good impression of the fine sense distinctions WordNet is capable of. The intended sense is chosen from considering both the context of the noun pair and what the most likely sense would be in general text.

Noun Pair - Car, Automobile

Ambiguous Word(s) : Car

Sense 1

car, auto, automobile, machine, motorcar, motor car
=> **motor vehicle, automotive vehicle**

Sense 2

car, gondola -- (carries personnel and cargo and power plant)
=> compartment

Sense 3

car, elevator car -- (where passengers ride up and down)
=> compartment

Sense 4

car, railway car, railroad car -- (adapted to the rails of railroad)
=> wheeled vehicle -- (moves on wheels)

Noun Pair - Gem, Jewel

Ambiguous Word(s) : Gem, Jewel

Sense 1

jewel, gem, precious stone
=> **jewelry, jewellery**

Sense 2

jewel, gem
=> precious stone

Sense 3

muffin, gem
=> quick bread

Sense 4

gem, treasure
=> art, fine art

Sense 5

gem, gemstone -- (a crystalline rock that can be cut and polished for jewelry)
=> crystal

Sense 1

jewel, gem
=> precious stone

Sense 2

jewel, gem, precious stone
=> **jewelry, jewellery**

Noun Pair - Journey, Voyage

Ambiguous Word(s) : Voyage

Sense 1

voyage -- (a journey to some distant place)

=> *journey -- (the act of traveling)*

Sense 2

ocean trip, voyage -- (an act of traveling by water)

=> water travel, travel by water

Noun Pair - Boy, Lad

Ambiguous Word(s) : Boy, Lad

Sense 1

male child, boy, child

=> *male, male person*

Sense 1

chap, fellow, lad, gent, fella

=> male. male person

Sense 2

boy -- (offensive name for Black man)

=> nigger, spade, coon, jigaboo, nigra

Sense 2

cub, lad, sonny, sonny boy

=> *male child, boy, child*

Sense 3

boy -- (a friendly informal reference to a grown man,

"he likes to play golf with the boys")

=> man, adult male -- (a grown man)

Noun Pair - Asylum, Madhouse

Ambiguous Word(s) : Asylum

Sense 1

mental hospital, mental institution, insane asylum, asylum -- (a hospital for mentally incompetent person)

=> *hospital, infirmary -- (where patients go for treatment)*

Sense 2

refuge, sanctuary, asylum

=> shelter

Noun Pair - Magician, Wizard

Ambiguous Word(s) : Magician, Wizard

Sense 1

sorcerer, magician, wizard, necromancer

=> *occultist*

Sense 1

sorcerer, magician, wizard, necromancer

=> *occultist*

Sense 2

magician, prestidigitator, conjurer, illusionist

=> performer, performing artist -

Sense 2

ace, adept, sensation, maven, virtuoso, genius, hotshot, star, whiz, whizz, wizard

=> expert -- (a person who performs skillfully)

Noun Pair - Stove, Furnace

Ambiguous Word(s) : Stove

Sense 1

stove, range, kitchen range, kitchen stove

=> *kitchen appliance*

Sense 2

stove -- (any heating apparatus)

=> heater, warmer -- (heats water or supplies warmth to a room)

Noun Pair - Food, Fruit

Sense 1

food, comestible, comestibles, edible, edibles, pabulum

=> *substance, matter*

Sense 2

food, nutrient

=> *substance, matter*

Ambiguous Word(s) : Food, Fruit

Sense 1

fruit -- (the ripened reproductive body of a seed plant)

=> *reproductive structure*

Sense 3

yield, fruit

=> *product, production*

Sense 2

fruit -- (edible part of a seed plant esp. one having sweet flesh)

=> *produce, green goods*

Noun Pair - Bird, Cock

Sense 1

bird

=> *vertebrate, craniate*

Sense 2

bird, fowl, poultry, volaille

=> *meat*

Sense 3

shuttlecock, bird, shuttle

=> *badminton equipment*

Sense 1

dame, doll, wench, skirt, cluck, bird

=> *girl, miss, gal, young lady, young woman, fille*

Ambiguous Word(s) : Bird, Cock

Sense 1

cock -- (adult male bird)

=> *bird*

Sense 2

stopcock, cock -- (device for regulating flow of a liquid)

=> *regulator*

Sense 3

cock, rooster -- (adult male chicken)

=> *chicken, Gallus gallus*

Sense 1

penis, phallus, member, cock, prick, dick, shaft, pecker, peter, tool

=> *erectile organ*

Noun Pair - Bird, Crane

Sense 1

bird

=> *vertebrate, craniate*

Sense 2

bird, fowl, poultry, volaille

=> *meat*

Sense 3

shuttlecock, bird, shuttle

=> *badminton equipment*

Sense 1

dame, doll, wench, skirt, chick, bird

=> *girl, miss, gal, young lady, young woman, fille*

Ambiguous Word(s) : Bird, Crane

Sense 1

crane

=> *wading bird, wader*

Sense 2

crane

=> *lifting device*

Noun Pair - Tool, Implement

Ambiguous Word(s) : Tool

Sense 1

creature, tool, puppet -- (a person who is used to perform unpleasant or dishonest tasks for someone else)

=> slave -- (a person who is owned by someone else)

Sense 2

penis, phallus, member, cock, prick, dick, shaft, pecker, peter, tool

=> erectile organ

Sense 3

tool -- (an implement used in the practice of a vocation)

=> **implement -- (a piece of equipment or tool used to effect an end)**

Sense 4

instrument, tool -- ("my greed was the instrument of my destruction")

=> means, way

Noun Pair - Monk, Brother

Ambiguous Word(s) : Brother

Sense 1

brother -- (a fellow member; usually of some religious group)

=> **member-- (one of the persons associated in a group) Sense 2**

Sense 2

brother, blood brother -- (a male with the same parents as someone else)

=> male sibling

Noun Pair - Implement, Crane

Ambiguous Word(s) : Crane

Sense 1

crane

=> **lifting device**

Sense 2

crane -- (large long-necked wading bird of marshes and plains in many parts of the world)

=> wading bird, wader -- (any of many long-legged birds that wade in water in search of food)

Noun Pair - Lad, Brother

Ambiguous Word(s) : Boy, Lad

Sense 1

cub, lad, sonny, sonny boy

=> **male child, boy, child**

Sense 2

chap, fellow, lad, gent, fella

=> male, male person

Sense 1

brother, blood brother

=> **male sibling**

Sense 2

brother -- (a fellow member; usually of some religious group)

=> **member**

Noun Pair - Journey, Car

Ambiguous Word(s) : Journey, Car

Sense 1

voyage -- (a journey to some distant place)
=> *journey -- (the act of traveling)*

Sense 2

ocean trip, voyage -- (an act of traveling by water)
=> *water travel, travel by water*

Sense 1

car, auto, automobile, machine, motorcar, motor car
=> *motor vehicle, automotive vehicle*

Sense 2

car, gondola
=> *compartment*

Sense 3

car, elevator car
=> *compartment*

Sense 4

car, railway car, railroad car
=> *wheeled vehicle -- (moves on wheels)*

Noun Pair - Food, Rooster

Ambiguous Word(s) : Food

Sense 1

food, comestible, comestibles, edible, edibles, pabulum
=> *substance, matter*

Sense 2

food, nutrient
=> *substance, matter*

Noun Pair - Coast, Hill

Ambiguous Word(s) : Hill

Sense 1

hill, hills
=> *natural elevation, elevation -- (a raised or elevated geological formation)*

Sense 2

mound, hill, pitcher's mound -- ((in baseball) the slight elevation on which the pitcher stands)
=> *sports equipment -- (equipment needed to participate in a particular sport)*

Noun Pair - Monk, Slave

Ambiguous Word(s) : Slave

Sense 1

slave, hard worker
=> *worker*

Sense 2

slave -- (a person who is owned by someone else)
=> *person, individual, someone, man, mortal, human, soul -- (a human being)*

Noun Pair - Lad, Wizard

Sense 1

cub, lad, sonny, sonny boy
=> *male child, boy, child*

Sense 2

chap, fellow, lad, gent, fella
=> male, male person

Ambiguous Word(s) : Lad, Wizard

Sense 1

sorcerer, magician, wizard, necromancer
=> *occultist*

Sense 2

ace, adept, sensation, maven, virtuoso, genius, hotshot,
star, whiz, whizz, wizard
=> expert -- (a person who performs skillfully)

Noun Pair - Chord, Smile

Sense 1

chord -- (a straight line connecting two points on a curve)
=> straight line -- ("the shortest distance between two points is a straight line")

Sense 2

chord -- (a combination of three or more notes that blend harmoniously when sounded together)
=> note, musical note, tone -- (a notation representing the pitch and duration of a musical sound)

Ambiguous Word(s) : Chord, Sense 1

Noun Pair - Glass, Magician

Sense 1

glass, drinking glass
=> *glassware, glasswork -- (articles made of glass)*

Sense 2

glass, glassful
=> containerful

Sense 3

looking glass, glass
=> mirror

Sense 4

glass -- (a brittle transparent solid)
=> solid

Sense 5

field glass, glass, spy glass
=> refracting telescope

Ambiguous Word(s) : Glass, Magician

Sense 1

sorcerer, magician, wizard, necromancer
=> *occultist*

Sense 2

magician, prestidigitator, conjurer, illusionist
=> performer, performing artist -

Noun Pair - Rooster, Voyage

Sense 1

voyage -- (a journey to some distant place)
=> *journey -- (the act of traveling)*

Sense 2

ocean trip, voyage -- (an act of traveling by water)
=> water travel, travel by water

Ambiguous Word(s) : Voyage

Noun Pair - Noon, String

Ambiguous Word(s) : String

Sense 1

string, twine -- (a lightweight cord)

=> *cord -- (a line made of twisted fibers or threads)*

Sense 2

succession, chain, string

=> *series -- (the relation between a number of events or things coming one after another)*

Sense 3

string, string of words, word string, linguistic string -- (a linear sequence of words as spoken or written)

=> *language, linguistic communication -- (a systematic means of communicating by the use of sounds or conventional symbols)*

Sense 4

string -- (a tightly stretched cord of wire or gut, which makes sound when plucked, struck, or bowed)

=> *cord -- (a line made of twisted fibers or threads)*

Sense 5

drawstring, string, tie -- (cord used as a fastener)

=> *cord -- (a line made of twisted fibers or threads)*

Sense 6

bowed stringed instrument, string -- ("the strings played superlatively well")

=> *stringed instrument*

Sense 7

chain, string, strand

=> *necklace*

Sense 8

string -- (a collection of objects threaded on a single strand)

=> *collection, aggregation, accumulation, assemblage -- (several things grouped together)*

Appendix D - Results of Psychological Evaluation

The tables in this appendix present the results evaluating the information content and conceptual distance similarity estimators against human judgements of similarity.

Noun Pair	Miller & Charles	Resnick	Information Based	Conceptual Distance
Car - Automobile	3.92	11.98	3.115	0.0
Gem - Jewel	3.84	18.34	4.853	0.0
Journey - Voyage	3.84	12.27	2.934	0.243
Boy - Lad	3.76	11.79	3.937	0.287
Coast - Shore	3.70	15.09	3.959	0.20
Asylum - Madhouse	3.61	20.08	5.298	0.125
Magician - Wizard	3.50	17.49	4.644	0.0
Midday - Noon	3.42	16.80	4.932	0.0
Stove - Furnace	3.11	5.90	0.978	0.156
Food - Fruit	3.08	5.47	2.144	1.867
Bird - Cock	3.05	13.06	3.224	0.787
Bird - Crane	2.97	13.06	3.224	0.626
Tool - Implement	2.95	9.96	2.269	0.296
Brother - Monk	2.82	5.74	0.973	2.252
Crane - Implement	1.68	5.74	1.348	1.257
Lad - Brother	1.66	5.90	0.973	2.411
Journey - Car	1.16	0.00	0.549	11.081
Monk - Oracle	1.10	5.74	0.973	2.186
Food - Rooster	0.89	4.65	0.549	6.768
Coast - Hill	0.87	10.72	2.272	1.221
Forest - Graveyard	0.84	0.00	0.999	9.721
Monk - Slave	0.55	5.74	0.973	1.731
Coast - Forest	0.42	0.00	0.549	9.906
Lad - Wizard	0.42	5.74	0.973	1.949
Chord - Smile	0.13	6.24	1.137	2.518
Glass - Magician	0.11	4.65	0.549	7.400
Noon - String	0.08	0.00	0.549	11.710
Rooster - Voyage	0.08	5.49	0.549	11.634

The columns of Table 1 represent the results returned by Miller and Charles' human subjects, Resnick's implementation of the information based similarity estimator, our

implementation, and the conceptual distance estimator using Sussna's weighting mechanism, respectively. Human subjects could rate similarity on a scale of 0 to 4, with 4 representing perfect synonymy and 0 representing very little similarity. It should be remembered that the higher the information content estimate the more similar, however, the lower the conceptual distance estimate, the more similar.

Table 2 shows the results of using a number of different weighting strategies for the conceptual distance similarity estimator. The test data set is the one used in [Mill91] and details of the different weighting formalisms can be found in section 4.3.

Noun Pair	Strength of Connotation	Local Density	Depth Scaling	Hybrid Depth Scaling
Car - Automobile	0.00	0.00	0.00	0.0
Gem - Jewel	0.00	0.00	0.00	0.0
Journey - Voyage	0.945	0.841	0.504	0.631
Boy - Lad	1.081	1.066	0.533	0.717
Coast - Shore	0.40	1.049	0.656	0.807
Asylum - Madhouse	0.25	0.536	0.357	0.429
Magician - Wizard	0.000	0.000	0.000	0.000
Midday - Noon	0.000	0.000	0.000	0.000
Stove - Furnace	0.210	0.210	0.210	0.210
Food - Fruit	3.159	3.291	2.446	2.804
Bird - Cock	2.993	2.835	2.255	2.543
Bird - Crane	2.352	2.241	1.614	2.070
Tool - Implement	1.165	1.086	0.905	0.988
Brother - Monk	8.754	8.695	4.726	6.019
Crane - Implement	4.787	4.680	3.548	3.932
Lad - Brother	8.607	9.798	5.690	7.150
Journey - Car	30.995	26.478	18.311	20.82
Monk - Oracle	7.564	7.813	3.140	4.795
Food - Rooster	25.428	21.982	12.395	15.734
Coast - Hill	4.344	4.701	2.829	3.526
Forest - Graveyard	26.050	22.292	10.423	13.343
Monk - Slave	6.856	6.876	3.423	4.467
Coast - Forest	26.297	22.977	12.128	15.213
Lad - Wizard	7.117	8.211	4.230	5.497
Chord - Smile	9.419	9.054	8.386	8.697
Glass - Magician	28.697	24.979	17.145	19.960
Noon - String	27.566	24.402	18.792	20.598
Rooster - Voyage	32.597	29.004	17.527	21.154

Appendix E - Syntactic Labels for RUCL Parser

This table list all the RUCL syntactic labels and shows how they are grouped into six categories.

Category	RUCL Label	Grammatical Description
Modifier	“<P”	Postposed complement of a prepositional
	“NN>“	Preposed noun modifier
	“NOM-OF”	Postposed of-expression
	“<NOM”	Postposed nominal complement
	“AD-A>”	Preposed modifier of an adjective
	“<P-FMAINV(n)”	Postposed verb complement of a prepositional
	“GN>“	Preposed genitive
	“<AD-A”	Postposed modifier of an adjective
Stopword	“DN>“	Preposed determiner
	“+FAUXV”	Finite auxiliary
	“CC”	Coordinating conjunction
	“QN>“	Preposed quantifier
	“INFMARK>“	Verb chain member
	“CS”	Subjunction
	“-FAUXV”	Nonfinite auxiliary
	“NEG”	Negation
	“DUMMY”	Non-label
Head	“SUBJ”	Subject
	“OBJ”	Object
	“APP”	Apposition
	“PCOMPL-S”	Subject predicate complement
	“N”	Title
	“PCOMPL-O”	Object predicate complement
	“NPHR”	Nominal Head
	“I-OBJ”	Indirect object
	“F-SUBJ”	Formal subject
Adverb	“ADVL ”	Adverb
	“O-ADVL ”	Adverb
Adjective	“AN>“	Adjective
Verb	“-FMAINV(n)”	Nonfinite verb
	“+FMAINV”	Finite main predicate
	“<NOM-FMAINV(n)”	Postposed verb as post modifier of a nominal

Appendix F - Statistics on WSJ Corpus

The test data set used in our document retrieval evaluation application are made up of a total of 173,256 articles from issues of the WSJ between the years 1986 and 1992 and a set of TREC queries. In this appendix there are examples of WSJ articles, TREC queries, as well as some statistics on the WSJ text corpus.

Two examples of the format of a WSJ article are as follows :

<DOC>
<DOCNO> WSJ870323-0178 </DOCNO>
<HL> Canadian Firms' New Orders</HL>
<DD> 03/23/87</DD>
<SO> WALL STREET JOURNAL (J)</SO>
<IN> CANADA </IN>
<DATELINE> OTTAWA </DATELINE>
<TEXT>

Canadian manufacturers' new orders fell to \$20.80 billion (Canadian) in January, down 4% from December's \$21.67 billion on a seasonally adjusted basis, Statistics Canada, a federal agency, said.

The decrease followed a 4.5% increase in December.

Manufacturers' shipments followed the same trend, falling 1.5% in January to \$21.08 billion, after a 2.8% increase the previous month.

The agency said there is "some indication of an upturn" in the recent irregular pattern of shipments, following the generally downward trend recorded during the first half of 1986.

</TEXT>
</DOC>

<DOC>
<DOCNO> WSJ870320-0065 </DOCNO>
<HL> Credit Ratings:
S&P Lowers Ratings
On Bethlehem Steel
Totaling \$761 Million</HL>
<DD> 03/20/87</DD>
<SO> WALL STREET JOURNAL (J)</SO>
<IN> BS
BOND MARKET NEWS (BON)
STOCK MARKET, OFFERINGS (STK) </IN>
<DATELINE> NEW YORK </DATELINE>
<TEXT>

Standard & Poor's Corp., citing concerns about Bethlehem Steel Corp.'s survival, downgraded its ratings on \$761 million of the steelmaker's debt.

The rating concern downgraded Bethlehem's senior debt to triple-C-plus from single-B-minus and its subordinated debt to triple-C-minus from triple-C.

S&P affirmed its single-C rating on Bethlehem preferred stock.

Dividends on the stock were suspended last year.

The actions reflect "concern for Bethlehem's viability over the intermediate term, rather than any immediate threat to solvency," S&P said.

In light of Bethlehem's "weak financial condition, the firm's ability to weather any substantial price competition or a general economic downturn is questionable", S&P said.

Choosing bankruptcy, "which relieved some of Bethlehem's competitors of their financial burdens", is still an alternative for the company "in the absence of any visible exit from the industry's morass", the rating concern said.

The steelmaker's liquidity, though, is strong relative to requirements this year, S&P said. Bethlehem's cash position improved to \$463 million at year's end from \$395 million Sept. 30 and \$99 million at the end of 1985.

But Bethlehem has raised its cash reserves through non-sustainable actions such as selling assets, reducing working capital and drawing on bank credit facilities, rather than from operations, S&P said.

A Bethlehem spokesman termed the downgrading "inappropriate" in light of the company's "stability" in the market, "realization of the anticipated improvements in our steel operations" and the current liquidity level. Bethlehem expects to maintain ample liquidity through this year, the spokesman said.

</TEXT>
</DOC>

The tags in these sample documents are as follows :

<DOC> - document begin	</DOC> - document end tag
<HL> - Document Title	<AUTHOR> - Author
<DATELINE> - place of writing	<DOCNO> - unique document ID
<TEXT> - start of natural language text	</TEXT> - end of text tag.

An example of the format of a TREC query is as follows :

<top>
<head> Tipster Topic Description
<num> Number: 101
<dom> Domain: Science and Technology
<title> Topic: Design of the "Star Wars" Anti-missile Defense System
<desc> Description:
Document will provide information on the proposed configuration, components, and technology of the U.S.'s "star wars" anti-missile defense system.

<narr> Narrative:
A relevant document will provide information which aids description of the design and technology to be used in the anti-missile defense system advocated by the Reagan administration, the Strategic Defense Initiative (SDI), also known as "star wars." Any reported changes to original design, or any research results which might lead to changes of constituent technologies,

are also relevant documents. However, reports on political debate over the SDI, or arms control negotiations which might encompass the SDI, are NOT relevant to the science and technology focus of this topic, unless they provide specific information on design and technology.

<con> Concept(s):

1. Strategic Defense Initiative, SDI, star wars, peace shield
2. kinetic energy weapon, kinetic kill, directed energy weapon, laser, particle beam, ERIS (exoatmospheric reentry-vehicle interceptor system), phased-array radar, microwave
3. anti-satellite (ASAT) weapon, spaced-based technology, strategic defense technologies

<fac> Factor(s):

<nat> Nationality: U.S.

</nat>

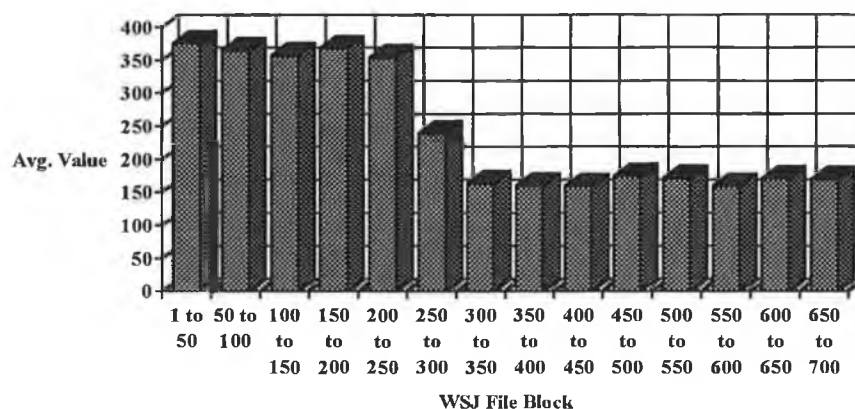
<def> Définition(s):

</top>

Note that the query is separated into fields and that the narrative field has a number of negative sentences. It should be noted that neither the semantic retrieval model developed in this research or the baseline pattern matching retrieval system process these statements. This is the original query from TREC, again for both retrieval systems an edited version of the query is sometimes used¹.

A number of small programs were written to gather a few statistics on the Wall Street Journal corpus. The corpus comes in the form of 710 one megabyte files and for the purposes of these statistics these files were grouped into fourteen 50 file blocks.

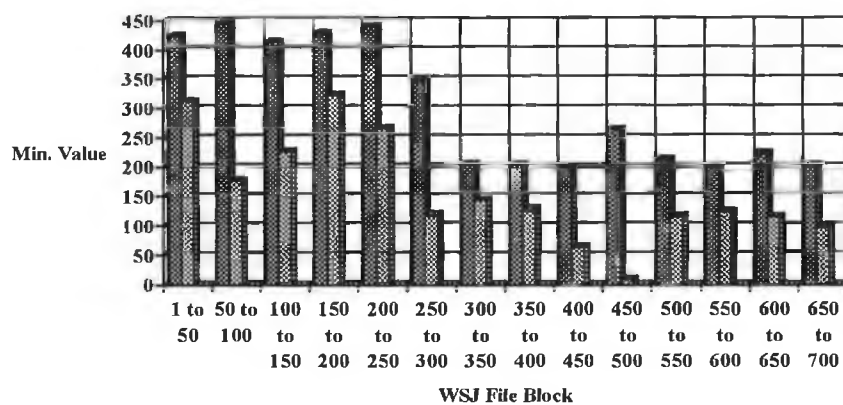
Average Number of Articles per WSJ file



¹ The organisers of TREC allow participants to modify queries as long as the meaning of the information request is not changed.

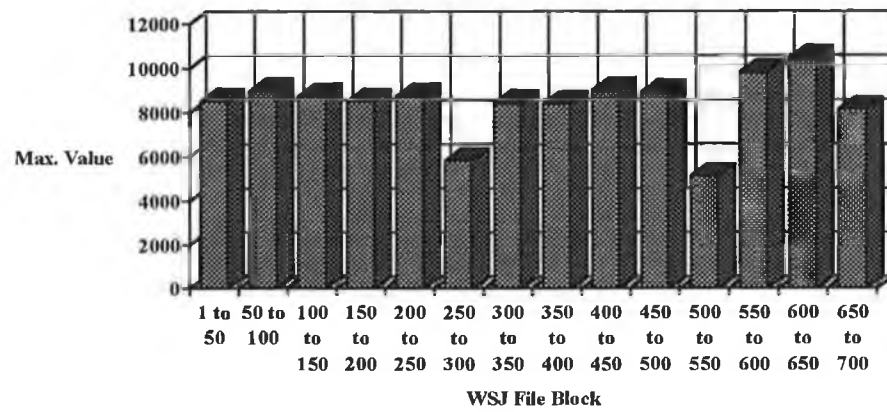
The first bar chart below shows the average number of articles per file for each of the fifty file blocks. As can be seen from the chart, the first block of 250 files have on average 350 articles per file, whereas the remaining files have an average of only 200 articles per file. This would suggest that articles in the last 276 megabytes are, on average, longer than those in the first batch. This is supported by the chart in figure 1, showing the average number of index terms per article. The articles in the first batch have an average of 370, or so, index terms, this is slightly less than the figure for the second batch, on average 420 terms per article. The second chart simply shows the maximum and minimum number of articles for each of the 50-file blocks.

Max and Min Number of Articles per WSJ file



As can be seen there is not a huge variability in the numbers of articles per file. This would be expected given the fact all WSJ files are roughly the same size, (around one megabyte). Finally, the third and fourth charts show the maximum and minimum number of index terms per article for each of the 50-file blocks. An index term is any word that is not a stop word, this is thus a good indicator of article lengths.

Max. Number of Index Terms per Article



It is clear from these charts that there is a huge variability in article lengths. They range from one liners to 15 - 20 pages in length. One of the outcomes of this finding was the decision to split articles into pages and to evaluate the relevance queries to articles on a page by page basis.

Appendix G - Stop list used in Preprocessor

Presented in this appendix is the list of words automatically stripped from the KB representation of an article. The list is made up terms which are proper nouns with acronym or slang interpretation, and terms occurring in Fox's general list of stop words.

Stop List			
1. Aba	40. February	80. Mere	119. Seven
2. Abc	41. Felt	81. Mil	120. Seventh
3. Ana	42. Fifth	82. Might	121. Show
4. Apr	43. First	83. Mol	122. Side
5. April	44. Five	84. Mon	123. Six
6. Area	46. Four	85. Monday	124. Sixth
7. Aug	47. Fourth	86. Month	125. Smith
8. August	48. Fri	87. Morning	126. Sol
9. Baa	49. Friday	88. Mrs	127. Someone
10. Back	50. Fun	89. Nine	128. Still
11. Behind	51. General	90. Ninth	129. Sun
12. Being	52. Goo	91. Nobody	130. Sunday
13. Ben	53. Good	92. Nothing	131. Ten
14. Bob	54. Group	93. Nov	132. Tenth
15. Boo	55. Have	94. November	133. Third
16. Case	56. High	95. Now	134. Three
17. Cent	57. iii	96. Nowhere	135. Thursday
18. Col	58. Important	97. Number	136. Tomorrow
19. Das	59. in.	98. Oct	137. Tues
20. Day	60. Ira	99. October	138. Tuesday
21. Deb	61. James	100. One	139. Turn
22. Dec	62. Jan	101. Open	140. Two
23. Decade	63. January	102. Order	141. Want
24. December	64. John	103. Pat	142. Way
25. Down	65. July	104. Place	143. Wed
26. Dun	66. Jun	105. Point	144. Wednesday
27. Ebs	67. June	106. Pre	145. Well
28. Eight	68. Just	107. Present	146. Wet
29. Eighth	69. Ken	108. Put	147. While
30. End	70. Kind	109. Right	148. Whole
31. Enough	71. Last	110. Ron	149. Wow
32. Eon	72. Less	111. Room	150. Wrong
33. Era	73. Look	112. Sat	151. xiv
34. Esp	74. Mac	113. Saturday	152. xxx
35. Est	75. Man	114. Sec	153. Year
36. Eta	76. Mar	115. Second	154. Young
37. Evening	77. March	116. Sep	
38. Fact	78. May	117. Sept	
39. Feb	79. Men	118. September	

Appendix H - Most Frequently Occurring WSJ Terms

This appendix shows the words that were found to appear in more than 10% of the articles in the Wall Street Journal. Most of these words are automatically removed from the list of index terms for a document because they are thought of as being non-content bearing. The numbers in brackets beside each word is a count of the number of articles the word appears in. In all we used 153,256 articles from the Wall Street Journal corpus.

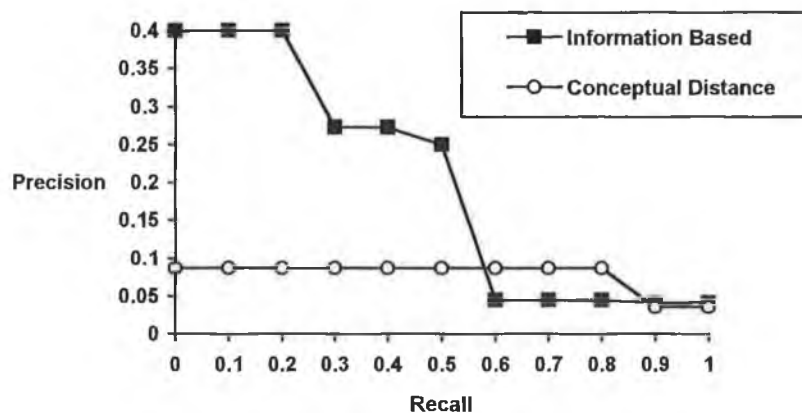
WSJ Word	Document Frequency	WSJ Word	Document Frequency
1. Agreement	22469	32. Meet	26778
2. Analyst	26106	33. Million	78871
3. Asset	118859	34. Month	49602
4. Average	18248	35. Move	29760
5. Billion	35756	36. Much	66568
6. Brief	21151	37. Nation	18559
7. Build	18763	38. Need	22600
8. Cent	27494	39. New York	41464
9. Comment	21114	40. Officer	26498
10. Concern	42470	41. Operation	24900
11. Deal	15883	42. Policy	16227
12. Development	15786	43. Position	16970
13. Dollar	21165	44. Raise	22256
14. Effort	15690	45. Reach	21775
15. Estimate	19991	46. Reporter	35564
16. Find	22586	47. Result	27820
17. Follow	23407	48. Return	17582
18. Fund	22479	49. Sale	45015
19. Gain	21729	50. See	31363
20. Give	36632	51. Share	57534
21. Help	28548	52. Shareholder	15492
22. Hold	41131	53. Spokesman	23632
23. Increase	42799	54. Start	20306
24. Interest	31949	55. Take	49199
25. Investor	24331	56. Thing	17914
26. Issue	35507	57. Use	42286
27. Keep	20317	58. Wall Street	43068
28. Look	21650	59. Week	39773
29. Make	66094	60. Yesterday	34863
30. Maker	23965		
31. Manager	16443		

Appendix I - Query break down of Thresholding Experiments

This appendix shows the queries used in the experiments carried out to determine values for the absolute noise threshold variables. Each query along with its KB representation is presented. The numbers to the left of the query terms represent the byte offsets of the corresponding KB synset. Interleaved between queries is an interpolated precision recall graph showing the performance of both the chosen configurations for information based and conceptual distance systems for that query. Values for the average non-interpolated precisions of both systems is also given.

<top> <head> Tipster Topic Description <num> Number: 101 <dom> Domain: Science and Technology <title> Topic: Design of the "Star Wars" Anti-missile Defense System <desc> Description: Document will provide information on the proposed configuration, components, and technology of the U.S.'s "star wars" anti-missile defense system. <narr> Narrative: A relevant document will provide information which aids description of the design and technology to be used in the anti-missile defense system advocated by the Reagan administration, the Strategic Defense Initiative (SDI), also known as "star wars." Any reported changes to original design, or any research results which might lead to changes of constituent technologies, are also relevant documents. However, reports on political debate over the SDI, or arms control negotiations which might encompass the SDI, are NOT relevant to the science and technology focus of this topic, unless they provide specific information on design and technology. <con> Concept(s): 1. Strategic Defense Initiative, SDI, star wars, pence shield 2. kinetic energy weapon, kinetic kill, directed energy weapon, laser, particle beam, IRIS (exoatmospheric reentry-vehicle interceptor system), phased-array radar, microwave 3. anti-satellite (ASAT) weapon, spaced-based technology, strategic defense technologies <fac> Factor(s): <nat> Nationality: U.S. </nat> <def> Definition(s): </top>	\$ST 1 defense_system 842957 technology 117398 research 44783 technology 173027 shield 783894 laser 1527855 laser_beam 102250 kinetic_energy 33510 weapon 842593 interception 354868 particle_beam 91998 microwave 73747 satellite 220657 space_age 224444 united_states 175569 missile 2 298428 star 220278 \$ET
---	--

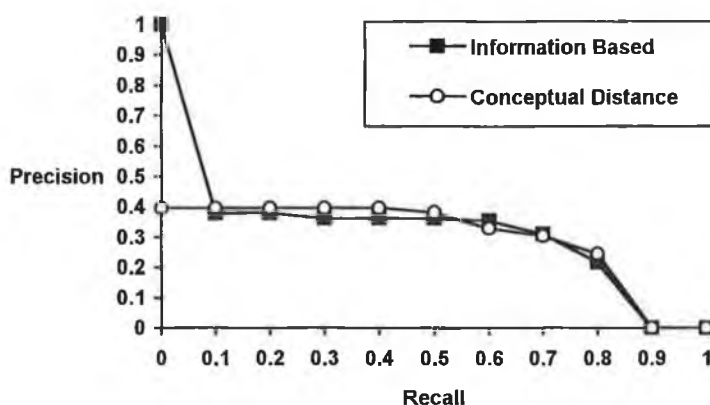
Query Number 1



Average Precision : Conceptual Distance (0.0622), Information Based (0.2857)

<top> <head> Tipster Topic Description <num> Number: 120 <dom> Domain: International Economics <title> Topic: Economic Impact of International Terrorism <desc> Description: Document will report on actual or alleged private sector economic consequences of international terrorism. <narr> Narrative: A relevant document will address the issue of how to calculate the private sector economic consequences of the activities of international terrorists. Reported consequences may be such information as corporate claims that a fall off in business resulted from customer fears over international terrorism, actual destruction of private property through terrorist acts, investments not made because of concerns over a terrorist presence in an area of potential investment, etc.. NOT relevant are costs attributable to government actions. <con> Concept(s): 1. terrorist, terrorism, international terrorist (or terrorism) <fac> Factor(s): <def> Definition(s): International Terrorism: as defined in Title 22 of the U.S. Code, Section 2656f(d), terrorism is premeditated, politically motivated violence perpetrated against noncombatant targets by subnational or clandestine agents, usually intended to influence an audience. By extension, international terrorism transcends national boundaries and involves the citizens, property, or territory of more than one country. </top>	\$ST 1 <table> <tr><td>terrorism</td><td>153392</td></tr> <tr><td>terrorist</td><td>1667106</td></tr> <tr><td>paramilitary</td><td>910868</td></tr> <tr><td>private_property</td><td>2846</td></tr> <tr><td>private_corporation</td><td>1006277</td></tr> <tr><td>bombing</td><td>186886</td></tr> <tr><td>bomb</td><td>1500918</td></tr> <tr><td>destruction</td><td>237608</td></tr> <tr><td>destructiveness</td><td>281585</td></tr> <tr><td>deterrent</td><td>84785</td></tr> <tr><td>intimidation</td><td>105035</td></tr> <tr><td>falloff</td><td>23495</td></tr> <tr><td>business</td><td>77509</td></tr> <tr><td>terrorism</td><td>100314</td></tr> <tr><td>investment</td><td>323332</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	terrorism	153392	terrorist	1667106	paramilitary	910868	private_property	2846	private_corporation	1006277	bombing	186886	bomb	1500918	destruction	237608	destructiveness	281585	deterrent	84785	intimidation	105035	falloff	23495	business	77509	terrorism	100314	investment	323332	\$ET	
terrorism	153392																																
terrorist	1667106																																
paramilitary	910868																																
private_property	2846																																
private_corporation	1006277																																
bombing	186886																																
bomb	1500918																																
destruction	237608																																
destructiveness	281585																																
deterrent	84785																																
intimidation	105035																																
falloff	23495																																
business	77509																																
terrorism	100314																																
investment	323332																																
\$ET																																	

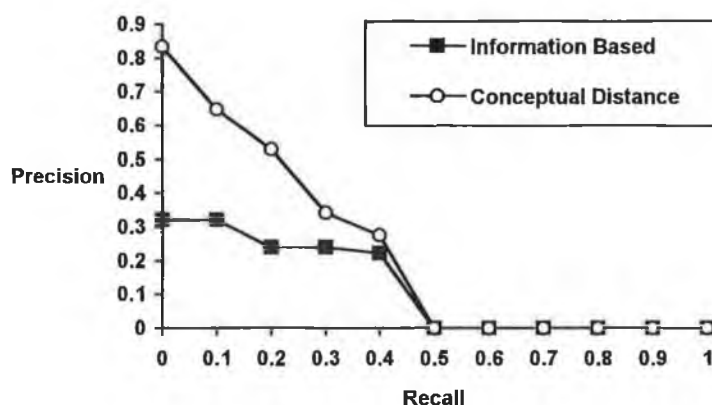
Query Number 2



Average Precision : Conceptual Distance (0.2741), Information Based (0.2861)

<top> <head> Tipster Topic Description <num> Number: 122 <dom> Domain: Medical & Biological <title> Topic: RDT&E of New Cancer Fighting Drugs <desc> Description: Document will report on the research, development, testing, and evaluation (RDT&E) of a new anti-cancer drug developed anywhere in the world. <narr> Narrative: A relevant document will report on any phase in the worldwide process of bringing new cancer fighting drugs to market, from conceptualization to government marketing approval. The laboratory or company responsible for the drug project, the specific type of cancer(s) which the drug is designed to counter, and the chemical/medical properties of the drug must be identified. <con> Concept(s): 1. cancer, leukemia 2. drug, chemotherapy <fac> Factor(s): <def> Definition(s): </top>	\$ST 2 <table> <tr> <td>cancer</td><td>172407</td></tr> <tr> <td>therapy</td><td>282756</td></tr> <tr> <td>drug</td><td>74131</td></tr> <tr> <td>research</td><td>44783</td></tr> <tr> <td>experimentation</td><td>189501</td></tr> <tr> <td>chemotherapy</td><td>398950</td></tr> <tr> <td>radiotherapy</td><td>398791</td></tr> <tr> <td>physical_property</td><td>53532</td></tr> <tr> <td>chemical</td><td>233308</td></tr> <tr> <td>leukemia</td><td>176108</td></tr> <tr> <td>approval</td><td>316704</td></tr> <tr> <td>carcinogen</td><td>67132</td></tr> </table> \$ET	cancer	172407	therapy	282756	drug	74131	research	44783	experimentation	189501	chemotherapy	398950	radiotherapy	398791	physical_property	53532	chemical	233308	leukemia	176108	approval	316704	carcinogen	67132
cancer	172407																								
therapy	282756																								
drug	74131																								
research	44783																								
experimentation	189501																								
chemotherapy	398950																								
radiotherapy	398791																								
physical_property	53532																								
chemical	233308																								
leukemia	176108																								
approval	316704																								
carcinogen	67132																								

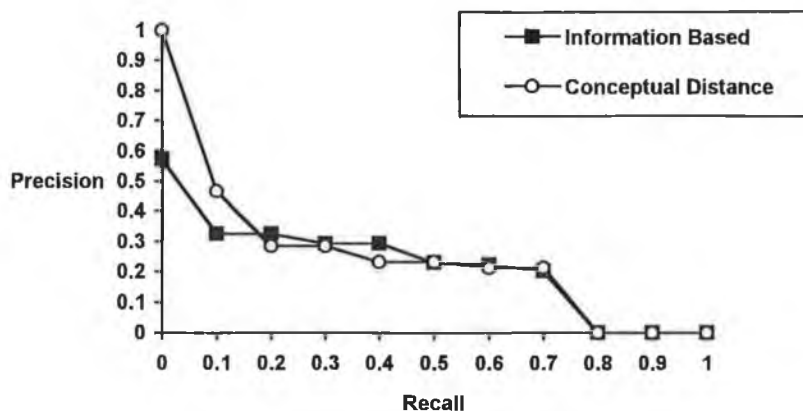
Query Number 3



Average Precision : Conceptual Distance (0.3140), Information Based (0.1744)

<top> <head> Tipster Topic Description <num> Number: 126 <dom> Domain: Medical & Biological <title> Topic: Medical Ethics and Modern Technology <desc> Description: Document will discuss ethical issues attendant to contemporary advances in medical technology. <narr> Narrative: A relevant document will report on non-traditional medical ethics issues which have been created by recent (post WWII) advances in medical science and/or the efforts of the medical community to deal with such ethical issues. Relevant reports include those on such issues as the decision process for selecting organ transplant recipients, whether to place or remove a patient from a life-support machine ("when to pull the plug"), whether or not to employ advanced techniques on defective newborns, and the use of fetal tissue and human cells in biomedical research. NOT relevant are reports on such traditional issues as abortion, mercy killing, and assisted suicide. Also NOT relevant are reports on the long-established legal/ethical issue of the right to privacy, as it relates to such contemporary public health problems as mandatory testing and contact tracing for AIDS. <con> Concept(s): 1. medical ethics 2. hospital ethicist, medical ethics counseling 3. triage, terminally ill, abortion, assisted suicide 4. fetal tissue, organ transplant, life support machine, artificial insemination <fac> Factor(s): <def> Definition(s): </top>	\$ST 3 <table> <tr><td>medicine</td><td>292094</td></tr> <tr><td>ethics</td><td>3864</td></tr> <tr><td>transplant</td><td>463735</td></tr> <tr><td>ethical_code</td><td>79647</td></tr> <tr><td>artificial_insemination</td><td>163852</td></tr> <tr><td>medical_community</td><td>964516</td></tr> <tr><td>fetus</td><td>44102</td></tr> <tr><td>tissue</td><td>212124</td></tr> <tr><td>abortion</td><td>238867</td></tr> <tr><td>organ</td><td>212452</td></tr> <tr><td>illness</td><td>78970</td></tr> <tr><td>hospital</td><td>1621191</td></tr> <tr><td>counsellor</td><td>977251</td></tr> <tr><td>code_of_conduct</td><td>453963</td></tr> <tr><td>termination</td><td>113262</td></tr> <tr><td>research</td><td>140956</td></tr> <tr><td>research</td><td>44783</td></tr> <tr><td>biomedical_science</td><td>215289</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	medicine	292094	ethics	3864	transplant	463735	ethical_code	79647	artificial_insemination	163852	medical_community	964516	fetus	44102	tissue	212124	abortion	238867	organ	212452	illness	78970	hospital	1621191	counsellor	977251	code_of_conduct	453963	termination	113262	research	140956	research	44783	biomedical_science	215289	\$ET	
medicine	292094																																						
ethics	3864																																						
transplant	463735																																						
ethical_code	79647																																						
artificial_insemination	163852																																						
medical_community	964516																																						
fetus	44102																																						
tissue	212124																																						
abortion	238867																																						
organ	212452																																						
illness	78970																																						
hospital	1621191																																						
counsellor	977251																																						
code_of_conduct	453963																																						
termination	113262																																						
research	140956																																						
research	44783																																						
biomedical_science	215289																																						
\$ET																																							

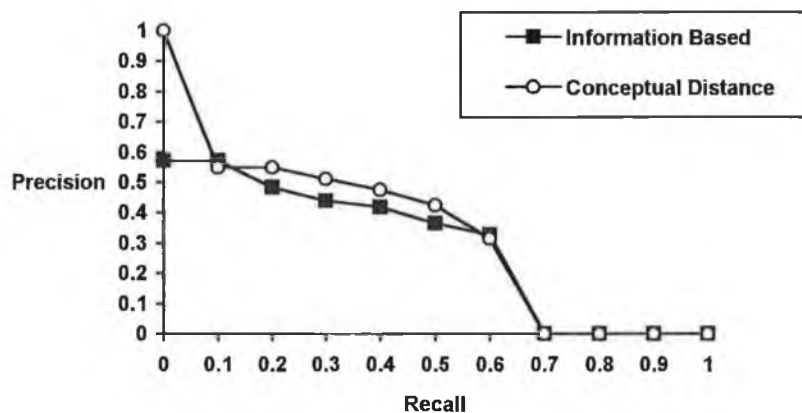
Query Number 4



Average Precision : Conceptual Distance (0.2456), Information Based (0.2022)

<top> <head> Tipster Topic Description <num> Number: 127 <dom> Domain: International Relations <title> Topic: U.S.-U.S.S.R. Arms Control Agreements <desc> Description: Document will report the terms of arms control agreements concluded, under negotiation, or proposed between the United States and the Soviet Union. <narr> Narrative: A relevant document must provide specific information on U.S.-U.S.S.R. arms control agreements, such as the weapons covered, size of reduction, verification procedures, etc. Debate in the U.S. Congress, or within Allied Nations or multilateral fora, is NOT relevant, unless such debates coincidentally provide new, concrete information on the terms of concluded, pending, or proposed arms control agreements. <con> Concept(s): 1. strategic arms limitation treaty, SALT, SALT II, START 2. intermediate range nuclear weapon, INF 3. medium-range nuclear missile, short-range nuclear weapon, nuclear testing, conventional arms 4. arms control, arms reduction, arms agreement, verification 5. Anti-ballistic Missile (ABM) Treaty of 1972 6. Arms Control and Disarmament Agency <fac> Factor(s): <nat> Nationality: U.S. <nat> Nationality: U.S.S.R. </fac> <def> Definition(s): </top>	\$ST 2 <table> <tr><td>arms_control</td><td>298939</td></tr> <tr><td>nuclear_weapon</td><td>1587781</td></tr> <tr><td>treaty</td><td>681911</td></tr> <tr><td>ballistic_missile</td><td>2840112</td></tr> <tr><td>arms_race</td><td>50060</td></tr> <tr><td>negotiation</td><td>614664</td></tr> <tr><td>united_states</td><td>175569</td></tr> <tr><td>soviet_union</td><td>117809</td></tr> <tr><td>moscow</td><td>221325</td></tr> <tr><td>agreement</td><td>471009</td></tr> <tr><td>limitation</td><td>157002</td></tr> <tr><td>nuclear_warhead</td><td>2297194</td></tr> <tr><td>intercontinental_ballistic_missile</td><td>3196905</td></tr> <tr><td>peace_treaty</td><td>761403</td></tr> <tr><td>diplomacy</td><td>715941</td></tr> <tr><td>cold_war</td><td>68116</td></tr> <tr><td>world_power</td><td>1016172</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	arms_control	298939	nuclear_weapon	1587781	treaty	681911	ballistic_missile	2840112	arms_race	50060	negotiation	614664	united_states	175569	soviet_union	117809	moscow	221325	agreement	471009	limitation	157002	nuclear_warhead	2297194	intercontinental_ballistic_missile	3196905	peace_treaty	761403	diplomacy	715941	cold_war	68116	world_power	1016172	\$ET	
arms_control	298939																																				
nuclear_weapon	1587781																																				
treaty	681911																																				
ballistic_missile	2840112																																				
arms_race	50060																																				
negotiation	614664																																				
united_states	175569																																				
soviet_union	117809																																				
moscow	221325																																				
agreement	471009																																				
limitation	157002																																				
nuclear_warhead	2297194																																				
intercontinental_ballistic_missile	3196905																																				
peace_treaty	761403																																				
diplomacy	715941																																				
cold_war	68116																																				
world_power	1016172																																				
\$ET																																					

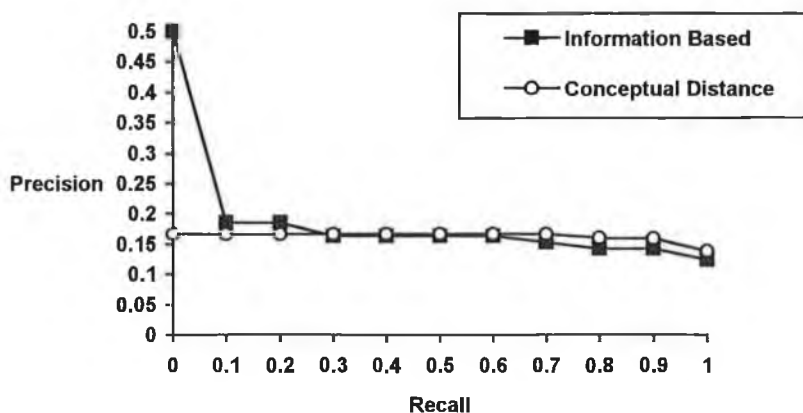
Query Number 5



Average Precision : Conceptual Distance (0.4559), Information Based (0.2640)

<top> <head> Tipster Topic Description <num> Number: 134 <dom> Domain: Medical & Biological <title> Topic: The Human Genome Project <desc> Description: Document will report on the objectives, processes, and organization of the human genome project. <narr> Narrative: A relevant document will report on the purposes, strategies, technologies, funding, and/or management of the human genome project. Although they may be scientifically pertinent, reports on general biomedical, genetic, and cell chemistry research, as well as related technologies, are NOT relevant to this topic unless they are explicitly linked to the human genome project. <con> Concept(s): 1. human genome, human genome initiative, human gene mapping 2. gene mapping, DNA sequencing 3. genetic marker, DNA segments <fac> Factor(s): <def> Definition(s): Human Genome Project: a largely U.S. funded "big science" project whose initial purpose is to map the nucleic acid sequences within the more than 50,000 genes carried on the 23 human chromosomes. Described differently, the project "is to produce a database listing the exact sequence of three billion base pairs along a single strand of human genetic material." </top>	<table> <tr> <td>\$ST 1</td><td></td></tr> <tr> <td>genetic_science</td><td>252815</td></tr> <tr> <td>nucleic_acid</td><td>731171</td></tr> <tr> <td>gene</td><td>720900</td></tr> <tr> <td>mapping</td><td>168158</td></tr> <tr> <td>gene-splicing</td><td>222651</td></tr> <tr> <td>genetic_constitution</td><td>147353</td></tr> <tr> <td>chromosome</td><td>216169</td></tr> <tr> <td>technology</td><td>173027</td></tr> <tr> <td>funding</td><td>323263</td></tr> <tr> <td>dna</td><td>1373981</td></tr> <tr> <td>research</td><td>140956</td></tr> <tr> <td>research</td><td>44783</td></tr> <tr> <td>researcher</td><td>443581</td></tr> <tr> <td>strategy</td><td>162071</td></tr> <tr> <td>human</td><td>3574</td></tr> <tr> <td>management</td><td>79932</td></tr> <tr> <td>\$ET</td><td></td></tr> </table>	\$ST 1		genetic_science	252815	nucleic_acid	731171	gene	720900	mapping	168158	gene-splicing	222651	genetic_constitution	147353	chromosome	216169	technology	173027	funding	323263	dna	1373981	research	140956	research	44783	researcher	443581	strategy	162071	human	3574	management	79932	\$ET	
\$ST 1																																					
genetic_science	252815																																				
nucleic_acid	731171																																				
gene	720900																																				
mapping	168158																																				
gene-splicing	222651																																				
genetic_constitution	147353																																				
chromosome	216169																																				
technology	173027																																				
funding	323263																																				
dna	1373981																																				
research	140956																																				
research	44783																																				
researcher	443581																																				
strategy	162071																																				
human	3574																																				
management	79932																																				
\$ET																																					

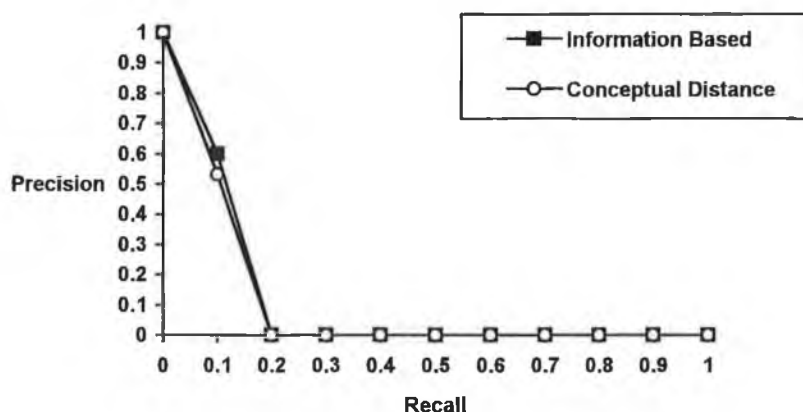
Query Number 6



Average Precision : Conceptual Distance (0.1304), Information Based (0.1304)

<top> <head> Tipster Topic Description <num> Number: 142 <dom> Domain: International Relations <title> Topic: Impact of Government Regulated Grain Farming on International Relations <desc> Description: Document will show how and why national governments supervise and protect their own growers of grain and/or how these policies impact a nation's foreign trade and diplomatic relations. <narr> Narrative: A relevant document will demonstrate specifically how a government anywhere in the world protects domestic production of grains (such as corn, rice, and wheat), provide the rationale for such policies (if available), and/or demonstrate the linkage between such domestic agricultural policies and a nation's foreign trade and diplomatic relations. <con> Concept(s): 1. export subsidies, export restitution, import quotas, farm-trade barriers 2. price supports, farm subsidies, agricultural supports 3. self-sufficiency, embargo, import dependence, high quality food supply, farm interests 4. grain, animal feed, corn, rice, wheat 5. foreign trade, trade negotiations, agricultural policy 6. EC's Common Agricultural Policy, Japan's Ministry of Agriculture, Forestry, and Fisheries, Zenchu, Liberal Democratic Party, LDP, U.S.'s Export Enhancement Program, food importing countries <fac> Factor(s): <def> Definition(s): </top>	\$ST 2 <table> <tr> <td>import_barrier</td><td>342627</td></tr> <tr> <td>agriculture</td><td>42734</td></tr> <tr> <td>price_support</td><td>68973</td></tr> <tr> <td>export_duty</td><td>68315</td></tr> <tr> <td>subsidy</td><td>60410</td></tr> <tr> <td>protective_tariff</td><td>68073</td></tr> <tr> <td>trade_embargo</td><td>495291</td></tr> <tr> <td>foreign_policy</td><td>230291</td></tr> <tr> <td>feed_grain</td><td>684210</td></tr> <tr> <td>corn</td><td>1297125</td></tr> <tr> <td>rice</td><td>1297846</td></tr> <tr> <td>wheat</td><td>2508144</td></tr> <tr> <td>farm</td><td>68912</td></tr> <tr> <td>domestic</td><td>509824</td></tr> <tr> <td>self-sufficiency</td><td>29767</td></tr> <tr> <td>restitution</td><td>52624</td></tr> <tr> <td>\$ET</td><td></td></tr> </table>	import_barrier	342627	agriculture	42734	price_support	68973	export_duty	68315	subsidy	60410	protective_tariff	68073	trade_embargo	495291	foreign_policy	230291	feed_grain	684210	corn	1297125	rice	1297846	wheat	2508144	farm	68912	domestic	509824	self-sufficiency	29767	restitution	52624	\$ET	
import_barrier	342627																																		
agriculture	42734																																		
price_support	68973																																		
export_duty	68315																																		
subsidy	60410																																		
protective_tariff	68073																																		
trade_embargo	495291																																		
foreign_policy	230291																																		
feed_grain	684210																																		
corn	1297125																																		
rice	1297846																																		
wheat	2508144																																		
farm	68912																																		
domestic	509824																																		
self-sufficiency	29767																																		
restitution	52624																																		
\$ET																																			

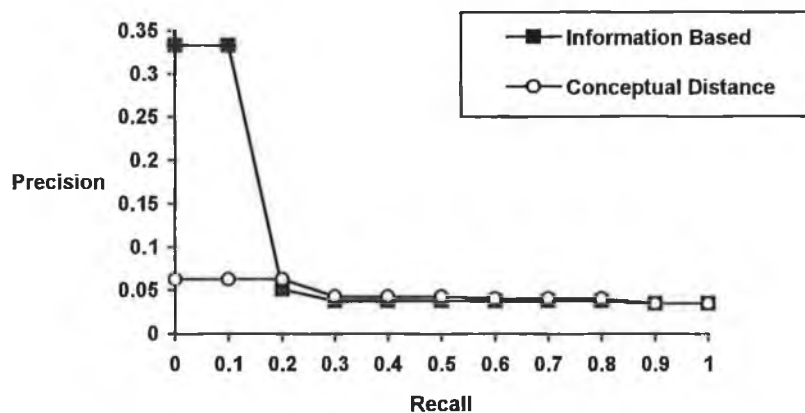
Query Number 7



Average Precision : Conceptual Distance (0.0819), Information Based (0.0819)

<top> <head> Tipster Topic Description <num> Number: 144 <dom> Domain: International Politics <title> Topic: Management Problems at the United Nations <desc> Description: Document will report on alleged corruption, incompetence, or inefficiency in the management of the United Nation's staff, activities, or specialized agencies. <narr> Narrative: A relevant document will provide information on management effectiveness by the United Nations, its specialized agencies, or in its running of such activities as peace keeping and relief operations. Allegations of management failings, as well as retorts to such charges, are relevant. However, charges of ideological bias or disputes over political and policy issues are NOT relevant. The focus is management effectiveness, or the lack thereof. <con> Concept(s): 1. United Nations, UN, Secretary General, United Nations Educational, Scientific, and Cultural Organization, UNESCO, Food and Agricultural Organization, FAO, World Health Organization, WHO, United Nations High Commissioner for Refugees, UNHCR, peace keeping 2. mismanagement, corruption, incompetence, inefficiency, ineffectiveness, overhead, extravagance, administrative costs 3. secretariat, international civil servants, bureaucracy 4. reform, budget controls, accounting, financial analysis, retrenchment <fac> Factor(s): <def> Definition(s): </top>	\$ST 0 <table> <tr><td>united_nations</td><td>962287</td></tr> <tr><td>bureaucracy</td><td>921292</td></tr> <tr><td>secretary_general</td><td>997978</td></tr> <tr><td>peace_corps</td><td>187744</td></tr> <tr><td>peacekeeper</td><td>1708387</td></tr> <tr><td>refugee</td><td>1015561</td></tr> <tr><td>disaster_area</td><td>14952</td></tr> <tr><td>educational_activity</td><td>11312</td></tr> <tr><td>assistance</td><td>67534</td></tr> <tr><td>corruption</td><td>256656</td></tr> <tr><td>incompetence</td><td>83030</td></tr> <tr><td>inefficiency</td><td>113342</td></tr> <tr><td>administration</td><td>196546</td></tr> <tr><td>budget</td><td>26274</td></tr> <tr><td>red_tape</td><td>330923</td></tr> <tr><td>accountability</td><td>109969</td></tr> <tr><td>management</td><td>79932</td></tr> </table> \$ET	united_nations	962287	bureaucracy	921292	secretary_general	997978	peace_corps	187744	peacekeeper	1708387	refugee	1015561	disaster_area	14952	educational_activity	11312	assistance	67534	corruption	256656	incompetence	83030	inefficiency	113342	administration	196546	budget	26274	red_tape	330923	accountability	109969	management	79932
united_nations	962287																																		
bureaucracy	921292																																		
secretary_general	997978																																		
peace_corps	187744																																		
peacekeeper	1708387																																		
refugee	1015561																																		
disaster_area	14952																																		
educational_activity	11312																																		
assistance	67534																																		
corruption	256656																																		
incompetence	83030																																		
inefficiency	113342																																		
administration	196546																																		
budget	26274																																		
red_tape	330923																																		
accountability	109969																																		
management	79932																																		

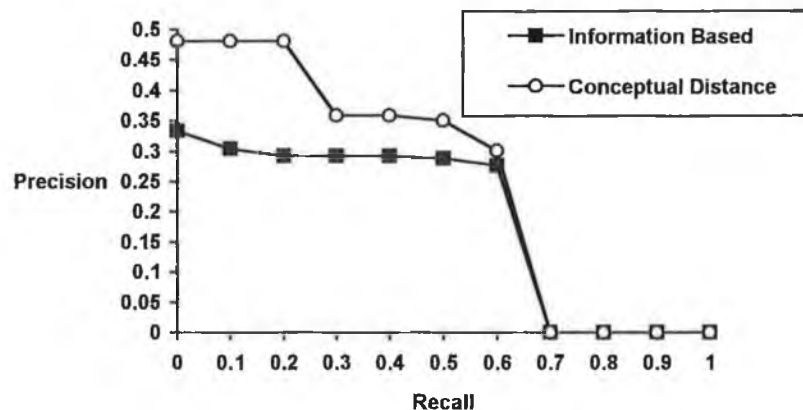
Query Number 8



Average Precision : Conceptual Distance (0.0413), Information Based (0.0788)

<top> <head> Tipster Topic Description <num> Number: 145 <dom> Domain: U.S. Politics <title> Topic: Influence of the "Pro-Israel Lobby" <desc> Description: Document will describe how, and how effectively, the so-called "pro-Israel lobby" operates in the United States. <narr> Narrative: A relevant document will identify and/or discuss the effectiveness of an organization, individual, or mechanism employed by the so-called "pro-Israel lobby" to advance the interests of the State of Israel within the United States and its government. Particularly useful would be a document which shows the lobby's impact on a specific policy decision by the U.S. government. <con> Concept(s): 1. zionism, American Jews, Jewish community, U.S. Jewish leaders 2. aid to Israel, military assistance, campaign contribution 3. U.S. arms sales to Egypt, Jordan, Saudi Arabia, or Kuwait 4. U.S. supporters of Israel, pro-Israel congressman or senator, pro-Israel lobbyist, Jewish lobby 5. American Jewish Congress, United Jewish Appeal, UJA, American Israel Public Affairs Committee, AIPAC, New Israel Fund, Committee for Economic Growth of Israel, Jewish Institute for National Security Affairs 6. Mossad, CIA, Israeli intelligence 7. dual loyalty, Jonathan Pollard, Pollard spy case, Rafael Eitan, Aviem Sella <fac> Factor(s): <nat> Nationality: U.S. and Israel </fac> <def> Definition(s): </top>	\$ST 0 <table> <tr><td>zionism</td><td>230486</td></tr> <tr><td>jew</td><td>515803</td></tr> <tr><td>united_states</td><td>175569</td></tr> <tr><td>israel</td><td>114893</td></tr> <tr><td>israeli</td><td>464666</td></tr> <tr><td>jordan</td><td>169872</td></tr> <tr><td>saudi_arabia</td><td>171598</td></tr> <tr><td>egypt</td><td>163649</td></tr> <tr><td>kuwait</td><td>170069</td></tr> <tr><td>public_relations</td><td>475024</td></tr> <tr><td>loyalty</td><td>410617</td></tr> <tr><td>lobbyist</td><td>392703</td></tr> <tr><td>sphere_of_influence</td><td>47102</td></tr> <tr><td>influence</td><td>64570</td></tr> <tr><td>u.s._government</td><td>944545</td></tr> <tr><td>political_party</td><td>188754</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	zionism	230486	jew	515803	united_states	175569	israel	114893	israeli	464666	jordan	169872	saudi_arabia	171598	egypt	163649	kuwait	170069	public_relations	475024	loyalty	410617	lobbyist	392703	sphere_of_influence	47102	influence	64570	u.s._government	944545	political_party	188754	\$ET	
zionism	230486																																		
jew	515803																																		
united_states	175569																																		
israel	114893																																		
israeli	464666																																		
jordan	169872																																		
saudi_arabia	171598																																		
egypt	163649																																		
kuwait	170069																																		
public_relations	475024																																		
loyalty	410617																																		
lobbyist	392703																																		
sphere_of_influence	47102																																		
influence	64570																																		
u.s._government	944545																																		
political_party	188754																																		
\$ET																																			

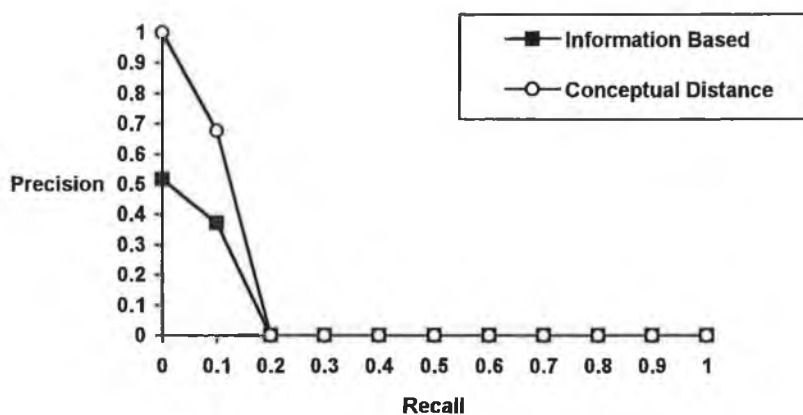
Query Number 9



Average Precision : Conceptual Distance (0.2295), Information Based (0.1690)

<top> <head> Tipster Topic Description <num> Number: 147 <dom> Domain: U.S. Economics <title> Topic: Productivity Trends in the U.S. Economy <desc> Description: Document will provide productivity statistics on the U.S. economy. <narr> Narrative: A relevant document will contain some macroeconomic datum useful in charting long-term productivity trends in the U.S. economy. Because useful in checking and calibrating aggregate figures, productivity statistics for specific sectors of the U.S. economy could be useful. However, statistics for individual enterprises, because not available in sufficient numbers, are NOT relevant. Also NOT relevant, unless containing confirmable productivity statistics, are reports on political discussions or press commentaries on the state of the U.S. economy. <con> Concept(s): 1. productivity plus United States 2. productivity plus trend, data, statistics, ratios, reports 3. productivity plus gains, growth, improvement, drop, decline, change 4. gross national product, GNP, growth rate, standard of living 5. recession, slump <fac> Factor(s): <nat> Nationality: U.S. </fac> <def> Definition(s): </top>	\$ST 0 <table> <tr><td>economic_condition</td><td>16333</td></tr> <tr><td>economist</td><td>990053</td></tr> <tr><td>trend</td><td>185678</td></tr> <tr><td>recession</td><td>100733</td></tr> <tr><td>slump</td><td>53593</td></tr> <tr><td>productivity</td><td>421461</td></tr> <tr><td>report</td><td>462977</td></tr> <tr><td>statistic</td><td>92985</td></tr> <tr><td>information</td><td>12657</td></tr> <tr><td>economy</td><td>18296</td></tr> <tr><td>production</td><td>2120</td></tr> <tr><td>progress_report</td><td>583310</td></tr> <tr><td>growth_rate</td><td>181421</td></tr> <tr><td>growth</td><td>200793</td></tr> <tr><td>gross_national_product</td><td>143589</td></tr> <tr><td>standard_of_living</td><td>19393</td></tr> </table> \$ET	economic_condition	16333	economist	990053	trend	185678	recession	100733	slump	53593	productivity	421461	report	462977	statistic	92985	information	12657	economy	18296	production	2120	progress_report	583310	growth_rate	181421	growth	200793	gross_national_product	143589	standard_of_living	19393
economic_condition	16333																																
economist	990053																																
trend	185678																																
recession	100733																																
slump	53593																																
productivity	421461																																
report	462977																																
statistic	92985																																
information	12657																																
economy	18296																																
production	2120																																
progress_report	583310																																
growth_rate	181421																																
growth	200793																																
gross_national_product	143589																																
standard_of_living	19393																																

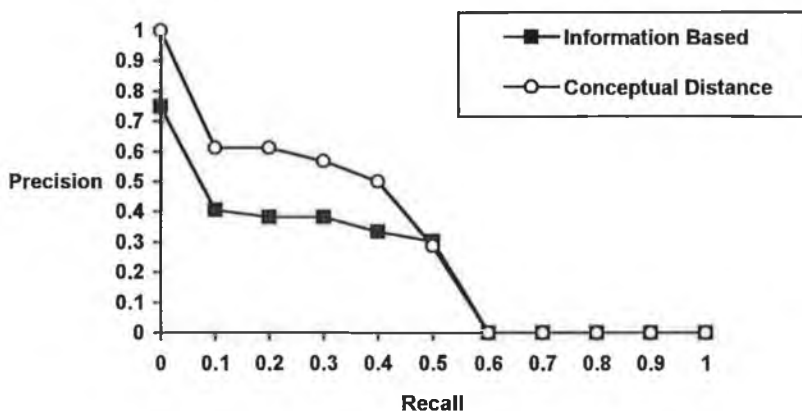
Query Number 10



Average Precision : Conceptual Distance (0.1098), Information Based (0.0743)

<top> <head> Tipster Topic Description <num> Number: 148 <dom> Domain: Military <title> Topic: Conflict in the Horn of Africa <desc> Description: Document will report on the Ethiopia-Somalia War, civil wars within those nations, and/or the movement of refugees fleeing armed conflicts between or within Ethiopia and Somalia. <narr> Narrative: A relevant document will provide information on the course of the Ethiopia-Somalia War (military and diplomatic developments), rebellions within Ethiopia, attempts to overthrow the Siad Barre regime in Somalia, and/or population movements within Ethiopia and Somalia, as well as refugee movements in and out of the neighboring states of Djibouti, Kenya, and Sudan. <con> Concept(s): 1. war, conflict 2. civil war, insurgency, coup, famine, guerrilla, refugee 3. Horn of Africa, Ethiopia, Somalia 4. Eritrea, Tigre (Tigray), Wollo (Wello), Addis Ababa 5. Kenya, Sudan, Djibouti 6. Eritrean People's Liberation Front, EPLF, Oromo Liberation Front 7. Mengistu, Siad Barre <fac> Factor(s): <def> Definition(s): </top>	\$ST 3 <table> <tr><td>ethiopia</td><td>163355</td></tr> <tr><td>somalia</td><td>166039</td></tr> <tr><td>war</td><td>73709</td></tr> <tr><td>civil_war</td><td>183812</td></tr> <tr><td>conflict</td><td>166262</td></tr> <tr><td>coup</td><td>20974</td></tr> <tr><td>famine</td><td>77255</td></tr> <tr><td>guerrilla</td><td>1707583</td></tr> <tr><td>rebellion</td><td>75925</td></tr> <tr><td>refugee</td><td>1015561</td></tr> <tr><td>kenya</td><td>163861</td></tr> <tr><td>sudan</td><td>166372</td></tr> <tr><td>djibouti</td><td>163172</td></tr> <tr><td>mogadiscio</td><td>221070</td></tr> <tr><td>addis_ababa</td><td>212049</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	ethiopia	163355	somalia	166039	war	73709	civil_war	183812	conflict	166262	coup	20974	famine	77255	guerrilla	1707583	rebellion	75925	refugee	1015561	kenya	163861	sudan	166372	djibouti	163172	mogadiscio	221070	addis_ababa	212049	\$ET	
ethiopia	163355																																
somalia	166039																																
war	73709																																
civil_war	183812																																
conflict	166262																																
coup	20974																																
famine	77255																																
guerrilla	1707583																																
rebellion	75925																																
refugee	1015561																																
kenya	163861																																
sudan	166372																																
djibouti	163172																																
mogadiscio	221070																																
addis_ababa	212049																																
\$ET																																	

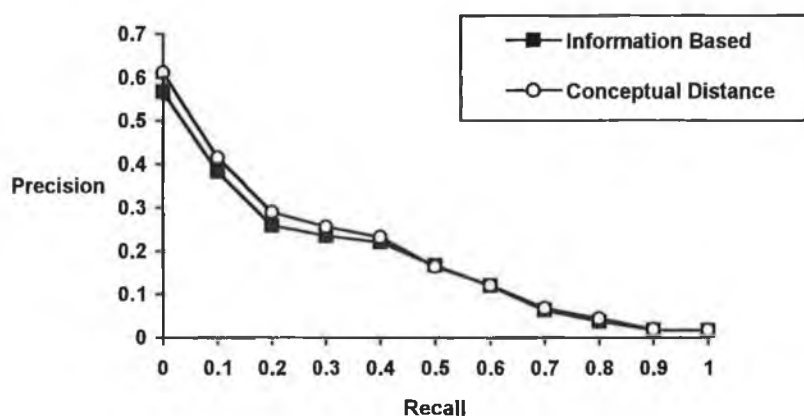
Query Number 11



Average Precision : Conceptual Distance (0.2618), Information Based (0.1956)

<top> <head> Tipster Topic Description <num> Number: 149 <dom> Domain: Law and Government <title> Topic: Industrial Espionage <desc> Description:	\$ST 6 espionage 469541 electronic_surveillance 282107 spy 1693825 commercial_bribery 465955 confidential_information 463186 trade_secret 582116 eavesdropper 476577 electronic_reconnaissance 470228 information_gathering 187796 intelligence_officer 1001405 intelligence_agent 1001405 intelligence_operation 321393 insider_trading 326550 theft 404617 document 452555 stock_market 10208 electronics_intelligence 469670 industry 950492 stock_exchange 102081 economy 18296 secret_agent 1001405 bribery 404334 security 57802 10 law_enforcement 194834 \$ET
Document will report on industrial espionage. <narr> Narrative: A relevant document will provide information on alleged or demonstrated acts of industrial espionage committed by agencies of any nation, by any corporation seeking information on a competitor through apparently illegal acts, or by private individuals seeking to collect and sell proprietary information. The espionage may be through any means -- electronic surveillance, bribery of employees, theft of documents, etc. -- whose purpose is to obtain technological or corporate secrets. Also relevant are the actions of governments to prevent the theft of economic secrets through legislation, regulation, or law enforcement. NOT relevant are such legally permissible actions as market research, hiring of consultants, and analysis of publicly available documents. <con> Concept(s): 1. industrial espionage, electronic surveillance, eavesdropping, electronic theft, inside information, market intelligence, industrial intelligence gathering, economic secrets, economic intelligence 2. industrial spy, disgruntled employee, electronic thief <fac> Factor(s): <def> Definition(s): </top>	

Query Number 12



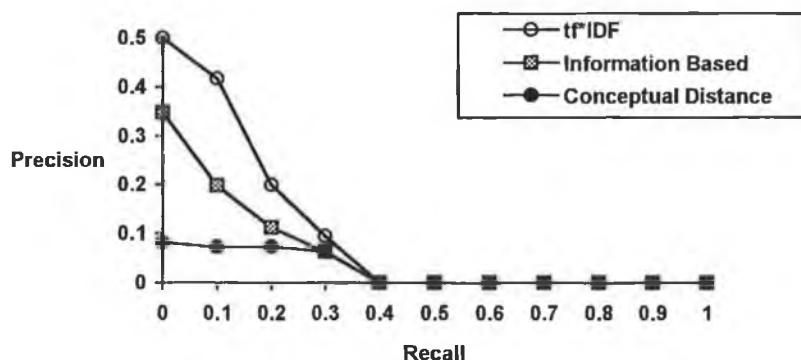
Average Precision : Conceptual Distance (0.3061), Information Based (0.3265)

Appendix J - Query break down of Full-test Experiments

This appendix shows the queries used in the evaluation experiments described in section 9.3. Each query along with it's KB representation is presented. The numbers to the left of the query terms represent the byte offsets of the corresponding KB synset. Interleaved between queries is an interpolated precision recall graph showing the performance of the information based, conceptual distance, and tf*IDF systems for that query. Again values for the average non-interpolated precisions for all three systems is also given.

<top> <head> Tipster Topic Description <num> Number: 106 <dom> Domain: Law and Government <title> Topic: U.S. Control of Insider Trading <desc> Description: Document will report proposed or enacted changes to U.S. laws and regulations designed to prevent insider trading. <narr> Narrative: A relevant document will contain information on proposed or enacted changes to U.S. laws and regulations, including state laws and stock market rules, which are aimed at increasing penalties or closing loopholes in existing institutional discouragements to insider trading. NOT relevant are reports on specific insider trading cases, such as the prosecutions and settlements related to the Boesky - Milken - Drexel Burnham Lambert scandal, unless the report also contains specific information on legal or regulatory change. <con> Concept(s): 1 insider trading 2 securities law, bill, legislation, regulation, rule 3 Insider Trading Sanctions Act, Insider Trading and Securities Fraud Enforcement Act 4 Securities and Exchange Commission, SEC, Commodity Futures Trading Commission, CFTC, National Association of Securities Dealers, NASD <fac> Factor(s): <nat> Nationality: U.S. </fac> <def> Definition(s): <top>	\$ST 3 insider_trading 326550 stock_market 102081 securities_law 57406 stock_trader 1674618 penalty 81267 regulation 584662 law 677365 united_states 175569 fraud 295955 fraudulent_scheme 154273 regulatory_authority 1015082 legislation 788163 legislation 194508 securities_industry 1008216 \$BT
---	---

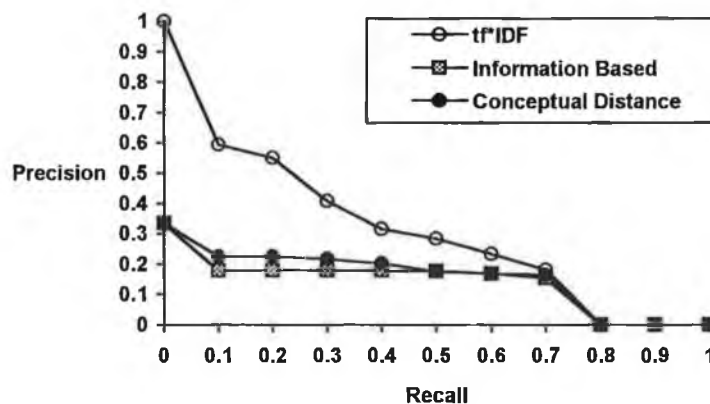
Query Number 1



Average Precision : tf*IDF (0.0876), Conceptual Distance (0.0221), Information Based (0.0508)

<top> <head> Tipster Topic Description <num> Number: 108 <dom> Domain: International Economics <title> Topic: Japanese Protectionist Measures <desc> Description: Document will report on Japanese policies or practices which help protect Japan's domestic market from foreign competition. <narr> Narrative: A relevant document will identify a Japanese law or regulation, a governmental policy or administrative procedure, a corporate custom, or a business practice which discourages, or even prevents, entry into the Japanese market by foreign goods and services. A document which reports generally on market penetration difficulties but which does not identify a specific Japanese barrier to trade is NOT relevant. <con> Concept(s): 1. Japan 2. Ministry of International Trade and Industry, MITI, Ministry of Foreign Affairs 3. protectionism, protect 4. tariff, subsidy, quota, dumping, obstruction, retaliation 5. structural impediment, product standard 6. trade dispute, barrier, tension, imbalance, practice 7. market access, free trade, liberalize, reciprocity 8. Super 301, 301 clause <fac> Factor(s): <nat> Nationality: Japan </fac> <def> Definition(s): </top>	\$ST 3 <table> <tr><td>japan</td><td>169675</td></tr> <tr><td>trade_barrier</td><td>342627</td></tr> <tr><td>trade_policy</td><td>262627</td></tr> <tr><td>protective_tariff</td><td>68073</td></tr> <tr><td>free_trade</td><td>189052</td></tr> <tr><td>government</td><td>186016</td></tr> <tr><td>policy</td><td>161600</td></tr> <tr><td>trade_embargo</td><td>49529</td></tr> <tr><td>tariff</td><td>58865</td></tr> <tr><td>subsidy</td><td>60410</td></tr> <tr><td>trade</td><td>66408</td></tr> <tr><td>balance_of_trade</td><td>189167</td></tr> <tr><td>self-sufficiency</td><td>29767</td></tr> <tr><td>market</td><td>44444</td></tr> <tr><td>japanese</td><td>465063</td></tr> <tr><td>foreign_minister</td><td>2358320</td></tr> <tr><td>foreign_policy</td><td>230291</td></tr> <tr><td>liberalization</td><td>454236</td></tr> <tr><td>tokyo</td><td>215353</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	japan	169675	trade_barrier	342627	trade_policy	262627	protective_tariff	68073	free_trade	189052	government	186016	policy	161600	trade_embargo	49529	tariff	58865	subsidy	60410	trade	66408	balance_of_trade	189167	self-sufficiency	29767	market	44444	japanese	465063	foreign_minister	2358320	foreign_policy	230291	liberalization	454236	tokyo	215353	\$ET	
japan	169675																																								
trade_barrier	342627																																								
trade_policy	262627																																								
protective_tariff	68073																																								
free_trade	189052																																								
government	186016																																								
policy	161600																																								
trade_embargo	49529																																								
tariff	58865																																								
subsidy	60410																																								
trade	66408																																								
balance_of_trade	189167																																								
self-sufficiency	29767																																								
market	44444																																								
japanese	465063																																								
foreign_minister	2358320																																								
foreign_policy	230291																																								
liberalization	454236																																								
tokyo	215353																																								
\$ET																																									

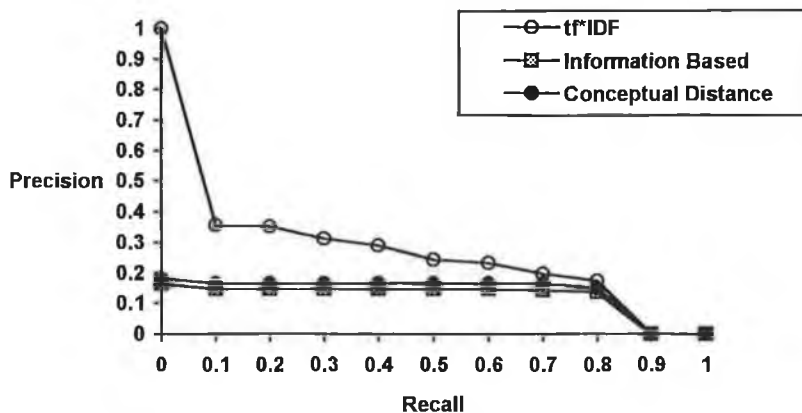
Query Number 2



Average Precision: tf*IDF (0.2797), Conceptual Distance (0.1394), Information Based (0.1212)

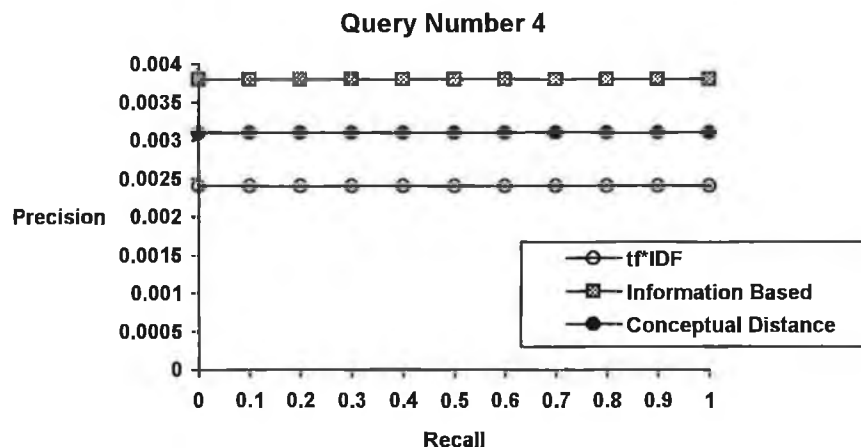
<top> <head> Tipster Topic Description <num> Number: 110 <dom> Domain: International Politics <title> Topic: Black Resistance Against the South African Government <desc> Description: Document will discuss efforts by the black majority in South Africa to overthrow domination by the white minority government. <narr> Narrative: A relevant document will discuss any effort by blacks to force political change in South Africa. The reported black challenge to apartheid may take any form -- military, political, or economic -- but of greatest interest would be information on reported activities by armed personnel linked to the African National Congress (ANC), either in South Africa or in bordering states. <con> Concept(s): 1. African National Congress, ANC, Nelson Mandela, Oliver Tambo 2. Chief Buthelezi, Inkatha, Zulu 3. terrorist, detainee, subversive, communist 4. Limpopo River, Angola, Botswana, Mozambique, Zambia 5. apartheid, black township, homelands, group areas act, emergency regulations <fac> Factor(s): <def> Definition(s): </top>	\$ST 4 <table> <tr><td>south_africa</td><td>166193</td></tr> <tr><td>apartheid</td><td>202157</td></tr> <tr><td>freedom_of_speech</td><td>249502</td></tr> <tr><td>racialism</td><td>198820</td></tr> <tr><td>limpopo</td><td>919435</td></tr> <tr><td>terrorist</td><td>1667106</td></tr> <tr><td>detainee</td><td>491699</td></tr> <tr><td>subversive</td><td>418268</td></tr> <tr><td>township</td><td>76840</td></tr> <tr><td>angola</td><td>161989</td></tr> <tr><td>african_nation</td><td>111371</td></tr> <tr><td>botswana</td><td>163555</td></tr> <tr><td>homeland</td><td>110388</td></tr> <tr><td>civil_rights_leader</td><td>418627</td></tr> <tr><td>negro</td><td>32927</td></tr> <tr><td>racism</td><td>177382</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	south_africa	166193	apartheid	202157	freedom_of_speech	249502	racialism	198820	limpopo	919435	terrorist	1667106	detainee	491699	subversive	418268	township	76840	angola	161989	african_nation	111371	botswana	163555	homeland	110388	civil_rights_leader	418627	negro	32927	racism	177382	\$ET	
south_africa	166193																																		
apartheid	202157																																		
freedom_of_speech	249502																																		
racialism	198820																																		
limpopo	919435																																		
terrorist	1667106																																		
detainee	491699																																		
subversive	418268																																		
township	76840																																		
angola	161989																																		
african_nation	111371																																		
botswana	163555																																		
homeland	110388																																		
civil_rights_leader	418627																																		
negro	32927																																		
racism	177382																																		
\$ET																																			

Query Number 3



Average Precision: tf*IDF (0.2339), Conceptual Distance (0.1193), Information Based (0.1116)

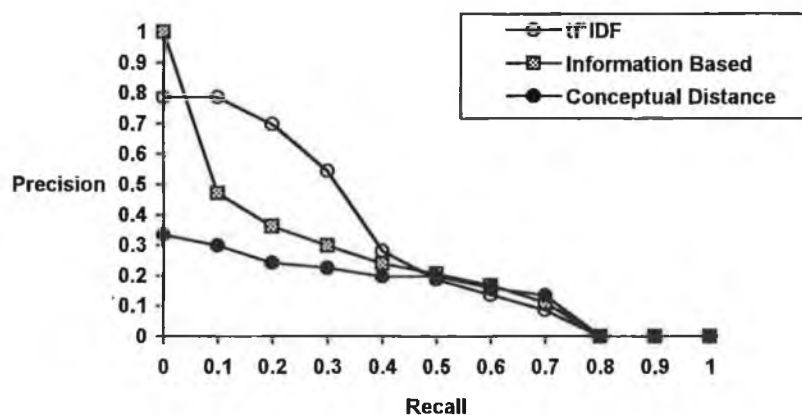
<top> <head> Tipster Topic Description <num> Number: 121 <dom> Domain: Medical & Biological <title> Topic: Death from Cancer <desc> Description: Document will discuss the life and death of a prominent U.S. person from a specific form of cancer. <narr> Narrative: A relevant document will provide obituary information on a prominent U.S. person who died of an identified type of cancer. In addition to the individual's name and cancer, the report must provide sufficient biographical information for a determination of why the life and contributions of the individual were worthy of some comment upon death. In other words, a one or two line obituary is NOT sufficient. <con> Concept(s): 1. cancer 2. death, obituary <fac> Factor(s): <nat> Nationality: U.S. <time> Time: current </fac> <def> Definition(s): </top>	\$ST 0 <table> <tr><td>cancer</td><td>172407</td></tr> <tr><td>human_death</td><td>34417</td></tr> <tr><td>obituary</td><td>594585</td></tr> <tr><td>tumor</td><td>127570</td></tr> <tr><td>illness</td><td>78970</td></tr> <tr><td>hospital</td><td>1 621191</td></tr> <tr><td>leukemia</td><td>176108</td></tr> <tr><td>celebrity</td><td>96606</td></tr> <tr><td>important_person</td><td>94531</td></tr> <tr><td>notable</td><td>930953</td></tr> <tr><td>lifetime</td><td>81907</td></tr> <tr><td>dead</td><td>6197</td></tr> <tr><td>lifework</td><td>144540</td></tr> <tr><td>life_story</td><td>698771</td></tr> </table> \$ET	cancer	172407	human_death	34417	obituary	594585	tumor	127570	illness	78970	hospital	1 621191	leukemia	176108	celebrity	96606	important_person	94531	notable	930953	lifetime	81907	dead	6197	lifework	144540	life_story	698771
cancer	172407																												
human_death	34417																												
obituary	594585																												
tumor	127570																												
illness	78970																												
hospital	1 621191																												
leukemia	176108																												
celebrity	96606																												
important_person	94531																												
notable	930953																												
lifetime	81907																												
dead	6197																												
lifework	144540																												
life_story	698771																												



Average Precision: tf*IDF (0.0021), Conceptual Distance (0.0026), Information Based (0.0030)

<top> <head> Tipster Topic Description <num> Number: 123 <dom> Domain: Medical & Biological <title> Topic: Research into & Control of Carcinogens <desc> Description: Document will report on studies into linkages between environmental factors or chemicals which might cause cancer, and/or it will report on governmental actions to identify, control, or limit exposure to those factors or chemicals which have been shown to be carcinogenic. <narr> Narrative: A relevant document will report on research into linkages between cancer and environmental hazards and/or the efforts of governments to limit exposure of their people to carcinogens. The governmental action may be of any category, e.g. entry into international agreements, enactment of domestic laws, issuance of administrative regulations, support of carcinogen research, air and soil sampling, launching of public education campaigns, etc. <con> Concept(s): 1. cancer, carcinogen 2. treaty, agreement, law, regulation, study, research, education, Super Fund <fac> Factor(s): <def> Definition(s): </top>	\$ST 2 <table> <tr><td>cancer</td><td>172407</td></tr> <tr><td>prevention</td><td>64426</td></tr> <tr><td>carcinogen</td><td>67132</td></tr> <tr><td>research</td><td>44783</td></tr> <tr><td>legal action</td><td>201129</td></tr> <tr><td>chemical</td><td>233308</td></tr> <tr><td>appraisal</td><td>100104</td></tr> <tr><td>administrative law</td><td>750635</td></tr> <tr><td>public exposure</td><td>87045</td></tr> <tr><td>jeopardy</td><td>57987</td></tr> <tr><td>government</td><td>186016</td></tr> <tr><td>identification</td><td>356078</td></tr> <tr><td>campaign</td><td>49582</td></tr> <tr><td>health</td><td>88690</td></tr> </table> \$ET	cancer	172407	prevention	64426	carcinogen	67132	research	44783	legal action	201129	chemical	233308	appraisal	100104	administrative law	750635	public exposure	87045	jeopardy	57987	government	186016	identification	356078	campaign	49582	health	88690
cancer	172407																												
prevention	64426																												
carcinogen	67132																												
research	44783																												
legal action	201129																												
chemical	233308																												
appraisal	100104																												
administrative law	750635																												
public exposure	87045																												
jeopardy	57987																												
government	186016																												
identification	356078																												
campaign	49582																												
health	88690																												

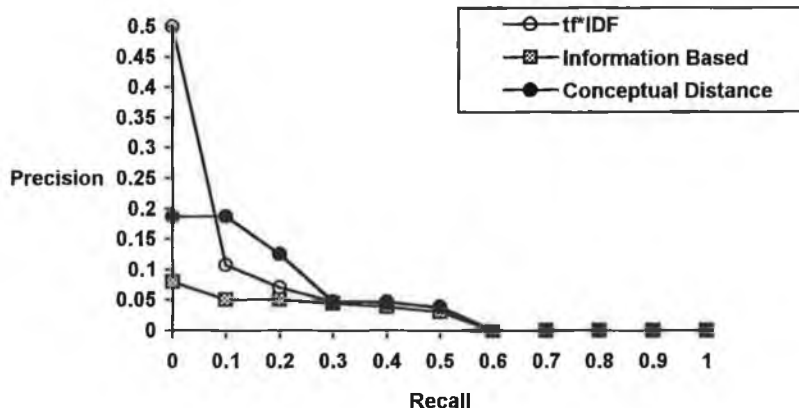
Query Number 5



Average Precision: tf*IDF (0.2888), Conceptual Distance (0.1460), Information Based (0.2197)

<top> <head> Tipster Topic Description <num> Number: 129 <dom> Domain: International Politics <title> Topic: Soviet Spying on the U.S. <desc> Description: Document will report spying by the USSR within U.S. territory or against U.S. interests overseas <narr> Narrative: A relevant document will discuss reported espionage by entities of the Soviet government - KGB, GRU, etc. - conducted within the territory of the United States of America, or against U.S. diplomatic or military facilities overseas. Reported entrapment or involvement of U.S. citizens, residents, or employees in Soviet spying, be it overseas or within U.S. territory, is also relevant. However, espionage cases involving states linked to the USSR - Czechoslovakia, Bulgaria, Cuba, etc. - are NOT relevant, unless linkage to Soviet intelligence can be demonstrated. <con> Concept(s): 1. USSR, U.S.S.R., Soviet, KGB, GRU, diplomat 2. spy, agent, spying, espionage, intelligence 3. snoop, bug, compromise, penetrate 4. counterintelligence, FBI, CIA, Pentagon, State Department <fac> Factor(s): <def> Definition(s): </top>	\$ST 2 <table> <tr><td>espionage</td><td>469541</td></tr> <tr><td>counterintelligence</td><td>416991</td></tr> <tr><td>spy</td><td>1693825</td></tr> <tr><td>soviet_union</td><td>117809</td></tr> <tr><td>communist</td><td>449921</td></tr> <tr><td>united_states</td><td>175569</td></tr> <tr><td>confidential_information</td><td>463186</td></tr> <tr><td>moscow</td><td>221325</td></tr> <tr><td>pentagon</td><td>964407</td></tr> <tr><td>security_intelligence</td><td>470369</td></tr> <tr><td>intelligence_agent</td><td>1001405</td></tr> <tr><td>intelligence_activity</td><td>321393</td></tr> <tr><td>bug</td><td>3157993</td></tr> <tr><td>state_department</td><td>1045465</td></tr> <tr><td>u.s._government</td><td>944545</td></tr> <tr><td>world_power</td><td>1016172</td></tr> <tr><td>cold_war</td><td>68116</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	espionage	469541	counterintelligence	416991	spy	1693825	soviet_union	117809	communist	449921	united_states	175569	confidential_information	463186	moscow	221325	pentagon	964407	security_intelligence	470369	intelligence_agent	1001405	intelligence_activity	321393	bug	3157993	state_department	1045465	u.s._government	944545	world_power	1016172	cold_war	68116	\$ET	
espionage	469541																																				
counterintelligence	416991																																				
spy	1693825																																				
soviet_union	117809																																				
communist	449921																																				
united_states	175569																																				
confidential_information	463186																																				
moscow	221325																																				
pentagon	964407																																				
security_intelligence	470369																																				
intelligence_agent	1001405																																				
intelligence_activity	321393																																				
bug	3157993																																				
state_department	1045465																																				
u.s._government	944545																																				
world_power	1016172																																				
cold_war	68116																																				
\$ET																																					

Query Number 6



Average Precision: tf*IDF (0.0564), Conceptual Distance (0.0453), Information Based (0.0218)

<top>
 <head> Tipster Topic Description
 <num> Number: 135
 <dom> Domain: Medical & Biological
 <title> Topic: Possible Contributions of Gene Mapping to Medicine
 <desc> Description:

Document will report on efforts to locate and describe genes linked to inherited human diseases and/or report on the potential medical contributions such information might yield.

<narr> Narrative:

A relevant document will report on any of the following: gene mapping research aimed at locating specific genes involved in those human diseases where genetic cause or predisposition have been implicated, attempts to describe the molecular structure and chemical defects of suspect genes (in either single gene or polygenic diseases), the development of diagnostic tests at the gene level, and research into possible treatments which might result from locating and analyzing genes involved in inherited human diseases. Reports on laboratory techniques or research projects which may aid the general fields of genetics and molecular biology, but which do not directly link the procedure or project to the study or treatment of a human genetic disease, are NOT relevant.

<con> Concept(s):

1. genetic disease, inherited disease, hereditary disease
2. gene mapping, DNA sequencing
3. gene therapy, gene test
4. chromosome analysis, DNA technology, genetic marker

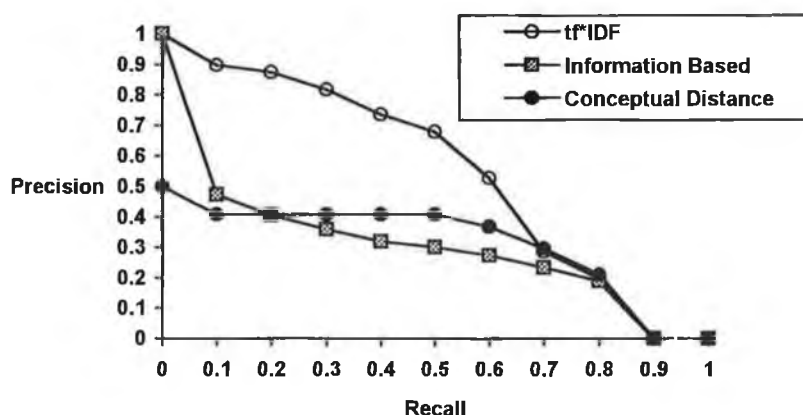
<fac> Factor(s):

<def> Definition(s):

</top>

\$ST 2
 genetic_disease 124060
 therapy 282756
 research 140956
 research 44783
 gene-splicing 222651
 genetic_science 252815
 genetic_endowment 50591
 gene 720900
 mapping 168158
 chromosome 216169
 dna 1373981
 \$ET

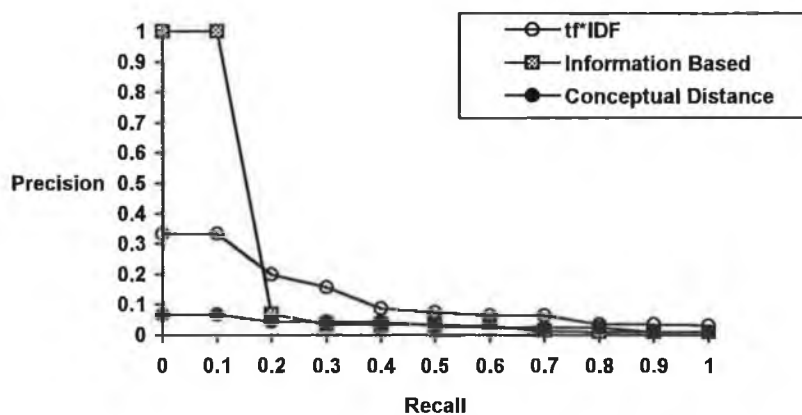
Query Number 7



Average Precision: tf*IDF (0.5506), Conceptual Distance (0.2877), Information Based (0.2745)

<top> <head> Tipster Topic Description <num> Number: 139 <dom> Domain: International Politics <title> Topic: Iran's Islamic Revolution - Domestic and Foreign Social Consequences <desc> Description: Document will report on the religious, legal, cultural, and social consequences of Iran's Islamic Revolution within Iran and abroad. <narr> Narrative: A relevant document will provide information which facilitates analysis of the non-political impact of Iran's Islamic Revolution on the people of Iran, as well as Muslims and others outside Iran. Relevant data on the religious, legal, cultural, and social dimensions of the Islamic Revolution should help describe how Ayatollah Khomeini's overthrow of the Shah impacted, and continues to impact, the lives of people. NOT relevant are reports focused on the human and economic losses associated with the Iran-Iraq War. <con> Concept(s): 1. Iran, Tehran, Qom, Ayatollah Khomeini 2. Islamic Revolution, Islamic Republic, Shiite, fundamentalism, Muslim, Moslem 3. women, chador, Islamic veil, headscarf 4. beard, village, hamlet 5. blasphemy, zealot, fanatic, repression, martyrdom, revolutionary guards 6. Salman Rushdie, Nazia Hassan, writer, artist, singer, television 7. NOT politics, NOT war, NOT elections, NOT international relations <fac> Factor(s): <def> Definition(s): </top>	\$ST 2 <table> <tr><td>iran</td><td>169343</td></tr> <tr><td>islamism</td><td>211971</td></tr> <tr><td>teheran</td><td>215140</td></tr> <tr><td>iranian</td><td>464450</td></tr> <tr><td>muslim</td><td>144990</td></tr> <tr><td>zealot</td><td>163562</td></tr> <tr><td>fundamentalism</td><td>270822</td></tr> <tr><td>turban</td><td>1416013</td></tr> <tr><td>social_psychology</td><td>220545</td></tr> <tr><td>sexism</td><td>198673</td></tr> <tr><td>social_system</td><td>18517</td></tr> <tr><td>acculturation</td><td>14951</td></tr> <tr><td>culture</td><td>135152</td></tr> <tr><td>religious_leader</td><td>146590</td></tr> <tr><td>\$ET</td><td></td></tr> </table>	iran	169343	islamism	211971	teheran	215140	iranian	464450	muslim	144990	zealot	163562	fundamentalism	270822	turban	1416013	social_psychology	220545	sexism	198673	social_system	18517	acculturation	14951	culture	135152	religious_leader	146590	\$ET	
iran	169343																														
islamism	211971																														
teheran	215140																														
iranian	464450																														
muslim	144990																														
zealot	163562																														
fundamentalism	270822																														
turban	1416013																														
social_psychology	220545																														
sexism	198673																														
social_system	18517																														
acculturation	14951																														
culture	135152																														
religious_leader	146590																														
\$ET																															

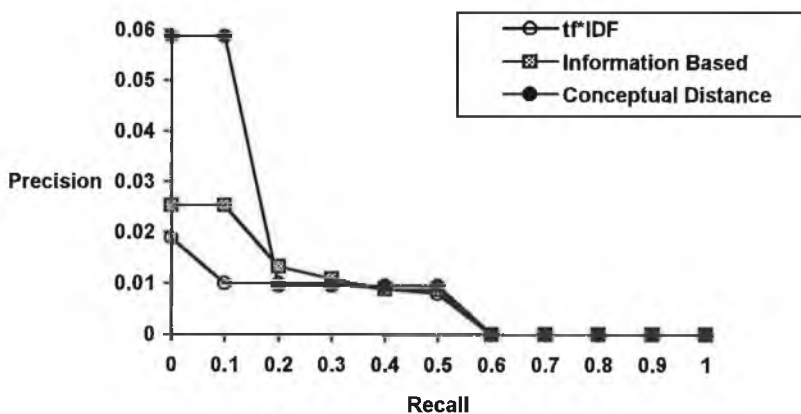
Query Number 8



Average Precision: tf*IDF (0.1084), Conceptual Distance (0.0319), Information Based (0.1243)

<top> <head> Tipster Topic Description <num> Number: 141 <dom> Domain: International Economics <title> Topic: Japan's Handling of its Trade Surplus with the U.S. <desc> Description: Document will report on Japanese efforts to deal with U.S. complaints regarding Japan's surplus in bilateral trade. <narr> Narrative: A relevant document will reveal Japanese government or government-inspired actions designed to reduce the gap or ameliorate frictions resulting from Japan's continuing surplus in bilateral trade with the U.S. The report must identify concrete actions which are traceable to the Japanese government. NOT relevant would be promises or proposals. Similarly, analyses of the U.S.-Japan trade relationship, trade negotiations, or political commentary thereon, are NOT relevant. Also, such economic data as foreign exchange and interest rate movements, corporate initiatives, or stock and bond market changes are NOT relevant, unless such information is explicitly linked to Japanese government efforts to deal with the bilateral trade surplus. <con> Concept(s): 1. Japan, Japanese government, Ministry of International Trade and Industry, MITI 2. U.S., Department of Commerce, U.S. Trade Representative 3. trade surplus, deficit, gap, imbalance, dispute 4. free trade, managed trade, barriers to trade, structural impediments <fac> Factor(s): <nat> Nationality: Japan and U.S. </fac> <def> Definition(s): </top>	\$ST 0 <table> <tr><td>japan</td><td>169675</td></tr> <tr><td>japanese</td><td>465063</td></tr> <tr><td>foreign_minister</td><td>2358320</td></tr> <tr><td>tokyo</td><td>215353</td></tr> <tr><td>national_trading_policy</td><td>262627</td></tr> <tr><td>balance_of_trade</td><td>189167</td></tr> <tr><td>government</td><td>186016</td></tr> <tr><td>u.s._government</td><td>944545</td></tr> <tr><td>united_states</td><td>175569</td></tr> <tr><td>measure</td><td>488643</td></tr> <tr><td>surplus</td><td>173468</td></tr> <tr><td>deficit_spending</td><td>327341</td></tr> <tr><td>department_of_commerce</td><td>1044041</td></tr> <tr><td>reduction</td><td>255671</td></tr> <tr><td>disparity</td><td>71577</td></tr> <tr><td>free_trade</td><td>189052</td></tr> <tr><td>trade_barrier</td><td>342627</td></tr> </table> \$ET	japan	169675	japanese	465063	foreign_minister	2358320	tokyo	215353	national_trading_policy	262627	balance_of_trade	189167	government	186016	u.s._government	944545	united_states	175569	measure	488643	surplus	173468	deficit_spending	327341	department_of_commerce	1044041	reduction	255671	disparity	71577	free_trade	189052	trade_barrier	342627
japan	169675																																		
japanese	465063																																		
foreign_minister	2358320																																		
tokyo	215353																																		
national_trading_policy	262627																																		
balance_of_trade	189167																																		
government	186016																																		
u.s._government	944545																																		
united_states	175569																																		
measure	488643																																		
surplus	173468																																		
deficit_spending	327341																																		
department_of_commerce	1044041																																		
reduction	255671																																		
disparity	71577																																		
free_trade	189052																																		
trade_barrier	342627																																		

Query Number 9



Average Precision: tf*IDF (0.0051), Conceptual Distance (0.0110), Information Based (0.0065)

<top>
 <head> Tipster Topic Description
 <num> Number: 143
 <dom> Domain: U.S. Politics
 <title> Topic: Why Protect U.S. Farmers?
 <desc> Description:

Document will contribute to an analysis of how and why, and at what cost, the U.S. federal government protects, supports, and controls U.S. farming.

<narr> Narrative:

A relevant document will describe actions taken by the U.S. government to protect U.S. farmers; or, reveal how farm price supports, export subsidies, import quotas, and other special farm policies impact the rural economy, as well as consumer costs and the federal budget; or, indicate a rationale used to justify farm support policies; or, show the relationship between agricultural policies and Congressional politics; or, suggest how national farm policies distort the "free" functioning of the agricultural sector of the national economy.

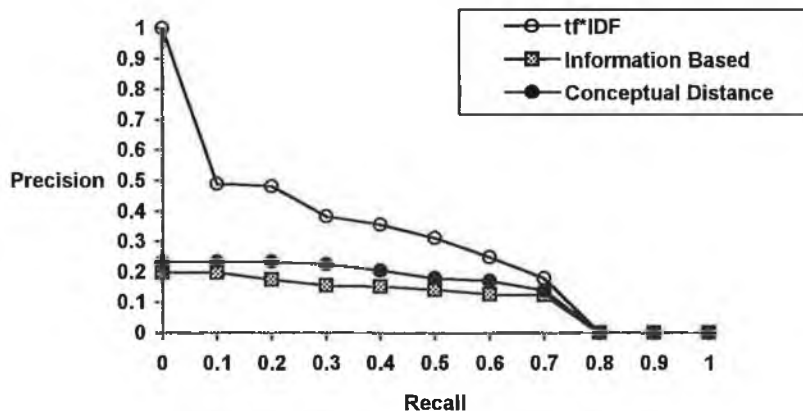
<con> Concept(s):

1. farm policy, farm exports, price supports, farm subsidy, deficiency payment, set-aside, supply controls, payment-in-kind, PIK, program crops, export subsidies, import quotas, crop base, production controls
2. Export Enhancement Program, EEP, Agriculture Adjustment Act, Targeted Export Assistance, TEA
3. rural districts, farm state, farming, farmers, agricultural sector
4. U.S. Department of Agriculture, Agriculture Department, USDA, U.S. Congress

<fac> Factor(s):
 <nat> Nationality: U.S.
 </fac>
 <def> Definition(s):
 </top>

\$ST 4
 price_support 68973
 agriculture 42734
 rural_area 46949
 subsidization 348906
 united_states 175569
 import_barrier 342627
 subsidy 60410
 restitution 52624
 protective_tariff 68073
 policy 185899
 crop_failure 88431
 farm 68912
 united_states_congress 917873
 crop 68916
 department_of_agriculture 1043864
 farming 69457
 farming_area 83777
 domestic 509824
 self-sufficiency 29767
 \$ET

Query Number 10



Average Precision: tf*IDF (0.2698), Conceptual Distance (0.1418), Information Based (0.1131)

<top>

<head> Tipster Topic Description

<num> Number: 146

<dom> Domain: International Relations

<title> Topic: Negotiating an End to the Nicaraguan Civil War

<desc> Description:

Document will report on the negotiating process leading to an end to the Nicaraguan civil war.

<narr> Narrative:

A relevant document will provide information on proposals for peace, negotiations on such proposals, terms of any agreements reached, problems in implementing agreements, or successful implementation of any agreement designed to facilitate an end to the civil war in Nicaragua.

<con> Concept(s):

1. Nicaragua, Costa Rica, El Salvador, Honduras, Guatemala
2. United States, Cuba, USSR, Soviet Union, Central America
3. Sandinista, Contra, Managua
4. peace talks, truce, cease-fire, elections, disarmament, nonlethal aid, turbista, La Prensa
5. Arias Plan, President Oscar Arias Sanchez, Oscar Arias
6. Daniel Ortega, Adolfo Calero, Aristides Sanchez, Violeta Chamorro, Cardinal Miguel Obando y Bravo, Cardinal Obando
7. Fidel Castro, Jose Napoleon Duarte, Alfredo Cristiani, Vinicio Cerezo, Jose Azcona

<fac> Factor(s):

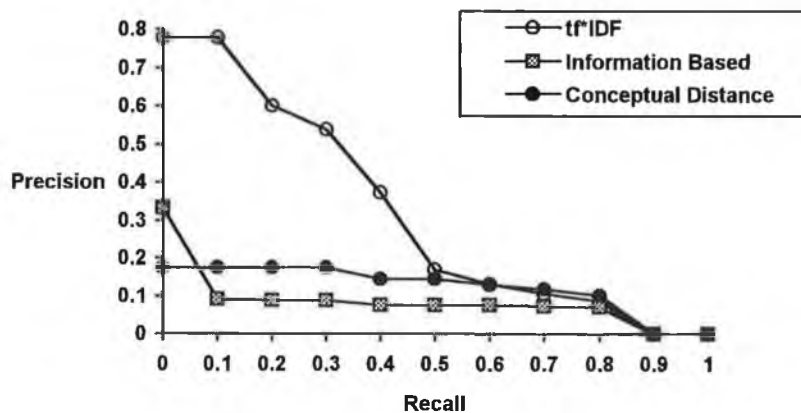
<def> Definition(s):

</top>

\$ST 2

nicaragua	206724
civil war	183812
nicaraguan	466631
managua	210091
costa rica	206256
guatemala	206352
cease-fire	66597
united states	175569
soviet union	117809
central america	49196
bloodshed	444729
truce	66597
disarmament	298939
diplomacy	71594
democracy	119844
\$ET	

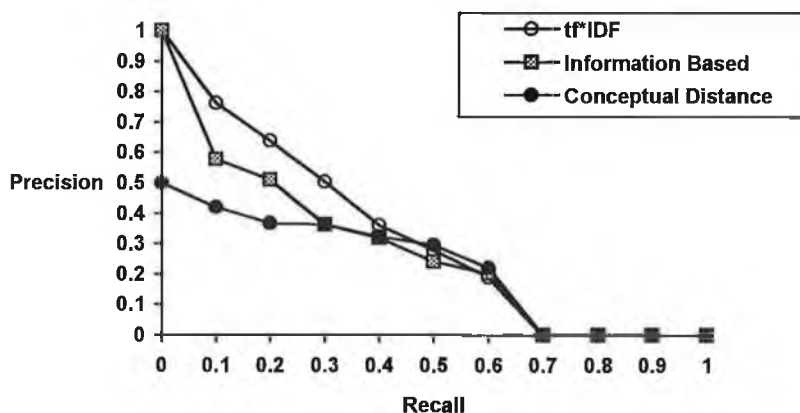
Query Number 11



Average Precision: tf*IDF (0.2870), Conceptual Distance (0.1109), Information Based (0.0735)

<top> <head> Tipster Topic Description <num> Number: 150 <dom> Domain: U.S. Politics <title> Topic: U.S. Political Campaign Financing <desc> Description: Document will report how U.S. politicians finance their election campaigns and/or moves to "reform" campaign finance practices. <narr> Narrative: A relevant document will show how U.S. politicians (federal, state, or local -- individually or as a group) pay for their election campaigns, the role played by "special interests" and contributors in the electoral process, allegations or evidence of campaign contributions buying political favors, and/or proposals to limit the cost of campaigns or "reform" electoral finance practices. <con> Concept(s): 1. campaign finance, campaign contribution, fund raising, political donation, honorarium, mother's milk of politics 2. campaign finance reform, public financing, ethics law 3. special interest, rich contributor, fat cat, lobbyist, political action committee, PAC 4. access-buying, political favor, electoral corruption <fac> Factor(s): <nat> Nationality: U.S. </fac> <def> Definition(s): </top>	\$ST 0 <table> <tr><td>political_campaign</td><td>156358</td></tr> <tr><td>campaign</td><td>50187</td></tr> <tr><td>campaign_for_governor</td><td>68258</td></tr> <tr><td>campaigner</td><td>449771</td></tr> <tr><td>contribution</td><td>322835</td></tr> <tr><td>contributor</td><td>983334</td></tr> <tr><td>donation</td><td>28363</td></tr> <tr><td>fund-raising_drive</td><td>155924</td></tr> <tr><td>honorarium</td><td>32586</td></tr> <tr><td>lobbyist</td><td>392703</td></tr> <tr><td>corruption</td><td>256656</td></tr> <tr><td>politician</td><td>131413</td></tr> <tr><td>politician</td><td>997783</td></tr> <tr><td>political_action_committee</td><td>1014136</td></tr> <tr><td>favor</td><td>195630</td></tr> <tr><td>reform</td><td>242563</td></tr> <tr><td>allegation</td><td>221800</td></tr> <tr><td>SET</td><td></td></tr> </table>	political_campaign	156358	campaign	50187	campaign_for_governor	68258	campaigner	449771	contribution	322835	contributor	983334	donation	28363	fund-raising_drive	155924	honorarium	32586	lobbyist	392703	corruption	256656	politician	131413	politician	997783	political_action_committee	1014136	favor	195630	reform	242563	allegation	221800	SET	
political_campaign	156358																																				
campaign	50187																																				
campaign_for_governor	68258																																				
campaigner	449771																																				
contribution	322835																																				
contributor	983334																																				
donation	28363																																				
fund-raising_drive	155924																																				
honorarium	32586																																				
lobbyist	392703																																				
corruption	256656																																				
politician	131413																																				
politician	997783																																				
political_action_committee	1014136																																				
favor	195630																																				
reform	242563																																				
allegation	221800																																				
SET																																					

Query Number 12

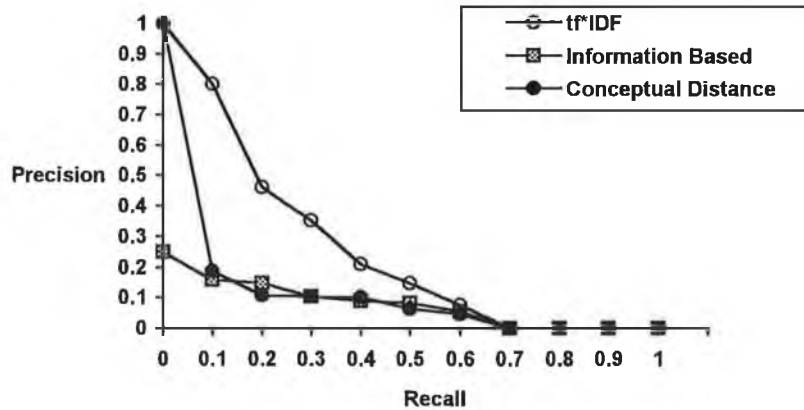


Average Precision: tf*IDF (0.2072), Conceptual Distance (0.1062), Information Based (0.1151)

Appendix K - Query break down of best queries

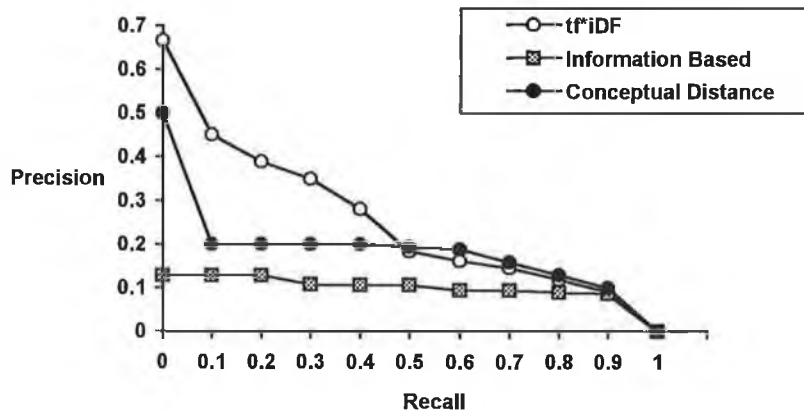
The graphs in this appendix present the results of running a full test evaluation on the best performing queries from the thresholding experiments, (queries 4, 5, 10, and 11). Refer to section 9.3.2 of chapter 9 for further details.

Query Number 4



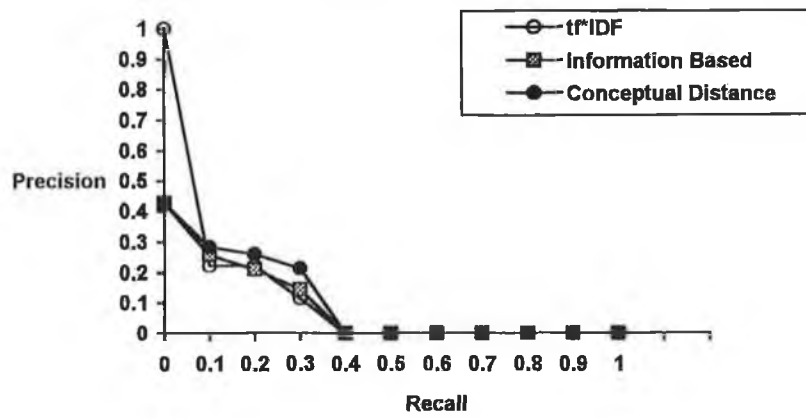
Average Precision: tf*IDF (0.2518), Conceptual Distance (0.0886), Information Based (0.0769)

Query Number 5



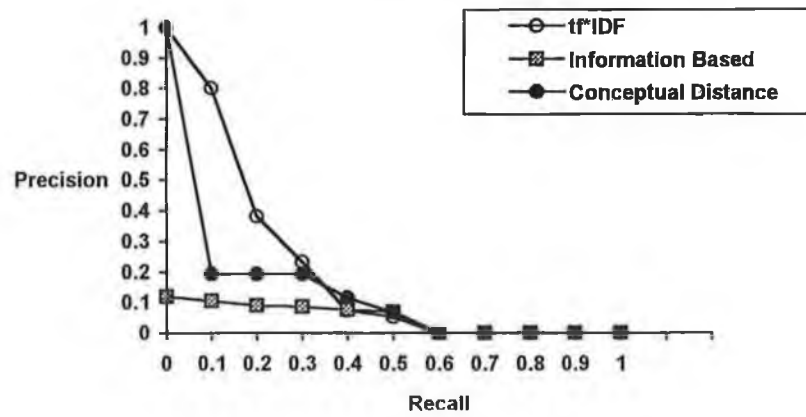
Average Precision: tf*IDF (0.2459), Conceptual Distance (0.1585), Information Based (0.0950)

Query Number 10



Average Precision: tf*IDF (0.0758), Conceptual Distance (0.0927), Information Based (0.0841)

Query Number 11



Average Precision: tf*IDF (0.1947), Conceptual Distance (0.0914), Information Based (0.0454)