CHAPTER 2

RELATED WORK

2.1 INTRODUCTION

Information Retrieval embraces the intellectual aspects of the description of information and its specification for search, including systems and techniques employed to carry out the retrieval operation (Mooers 1951). Information Retrieval needs to efficiently and effectively fetch relevant documents from a large archive in response to user’s query and rank these documents according to relevance. In this scenario there is a need for effective matching of documents to suit users’ information needs and appropriate representation of the documents to aid this matching process. Documents are unstructured natural language texts and their representation can vary from simple collection of words approach to complex rich semantic representation of the meaning of the documents. Richer representations call for computationally expensive techniques but result in more effective retrieval.

IR approaches are broadly classified as Statistical and Linguistic oriented though in most approaches there is an integration of the two. Statistical approaches are more concerned with the matching and ranking process while linguistic approaches are more concerned with building richer document representations; however both approaches need to address both the representation and matching issues. Classification of IR approaches can be done
from different perspectives. One perspective is to classify them based on the complexity of matching techniques used to select the documents. The other and more common perspective is to classify IR systems based on both the richness of the representation of the documents and the associated matching process. From this viewpoint IR systems can be classified as Boolean models, Statistical models, Linguistic Models and Logical Models (Figure 2.1). The analysis complexity varies from simple frequency count to complex semantic analysis of the documents. In this thesis an integrated statistical and linguistic oriented approach to build multidimensional document representation has been attempted.

![Figure 2.1 Classification of IR models](image)

Boolean models consider only terms (which may be words or phrases) whereas the statistical model considers the importance of terms in the
documents by assigning weights. The linguistic model analyses documents from the morphological, lexical, syntactic, semantic and/or contextual perspectives. Logical Models try to represent documents that convey uncertainty, user's feedback or semantics of the document. The use of Logic based models to represent uncertainty and terminological knowledge associated with the document can lead to a limited extent of inference from the document. Boolean and statistical models may use a limited level of linguistic processing (eg. Parts of speech assignment, morphological stemming etc) and Linguistic models invariably use statistical approaches for filtering purposes. In the following sections the various models have been explained in detail.

2.2 BOOLEAN MODELS

Boolean models can be based on Classical Boolean operators or can be refined with proximity operators as in Smart Boolean models. In order to overcome the all-or-nothing disadvantage of classical approaches, extended Boolean models have been proposed that soften the classical Boolean operators by associating weights to the matching process and allowing ranking of retrieved documents.

2.2.1 Classical Boolean Models

The Classical Boolean model is the simplest of the models and relies on the use of classical Boolean operators. In this approach, the query is formulated as a Boolean combination of terms. A conventional Boolean query uses the classical operators AND, OR and NOT (Heaps 1978) having great expressive power and clarity. However a classical Boolean query returns either
true or false. In other words, a document is either relevant to the query or irrelevant to it and hence ranking a document is not possible. This is one of the greatest disadvantages of Classical Boolean models.

Boolean models for IR offer a multitude of techniques to refine the query such as imposing proximity constraints, specifying or restricting the topic of search and varying the degree of stemming, restricting the search to the title and performing exact rather than truncated word matches. An additional proximity operator may be added to the classical set to narrow down the search. The proximity operator indicates the extent of proximity the terms have to be in the text, to satisfy the query condition. The proximity can be specified as words, phrases, sentences or any other unit of a text, or by an integer. The proximity value can vary from zero, which indicates terms have to be adjacent to n, which indicated terms have to be n terms apart. The proximity operator can also specify the order of the terms. Based on the proximity operator the Smart Boolean approach (Marcus 1994) offers a set of strategies for modifying a query based on the received relevance feedback or the expressed need to narrow or broaden the query.

The Classical Boolean operator is an approximate way of expressing phrase and thesaurus relations where the AND operator groups terms into phrases whereas the ORed terms are in a sense considered equivalent. The logical structure of the Boolean model allows preciseness in query expression; however this very property also leads to the problem of proper human expression of queries especially for naïve users. The interpretation of the Boolean AND and OR is different from the human interpretation of the same terms. (Cooper 1988) This leads to the design of methods for automatic
generation of queries. The Boolean expressions are constructed by ORing each query term with any stored synonyms and then ANDing these clusters together (Anick et al 1994). In Anick’s work the query terms after stemming and removing stop words, are grouped together using AND operators. Each term is expanded or ORed with any synonyms which are taken from an on-line thesaurus. On the other hand, Salton expresses the Boolean expressions with the terms given by the user rather than synonyms taken from the thesaurus. Salton looked for pairs and triples of these user-supplied terms that co-occur in one or more documents. Salton (Salton 1974) has also introduced ranking of documents which in a very limited way allows the user to control the number of documents retrieved but does not contribute to relevance ranking.

Boolean retrieval is more effective in the later stages of the search process, because of the clarity and exactness with which relationships between concepts can be represented. However Boolean queries are difficult to construct effectively because they resort to users’ knowledge of English and their corresponding mapping to Boolean queries. Boolean models allow only exact matches and hence if the query is too general a large number of documents are retrieved and if the query is strict very few documents are retrieved. However the most important disadvantage of classical Boolean models is the absence of the relevance ranking of the retrieved documents.

2.2.2 Extended Boolean Models

In spite of the addition of the proximity operator, Boolean conditions either retrieve large number of documents or none at all. To solve this problem and to provide relevance ranking, the extended Boolean model was designed
with an assignment of weights to the query or document terms. This model softens the Boolean operators by allowing approximate matches of the documents that satisfy the Boolean query and allowing closeness of matching to indicate ranking.

Expanded term weighting operations make ranking of documents possible, where the terms in the document could be weighted according to their frequency in the document (Salton 1983). A number of extended Boolean models have been developed (Paice 1984, Waller et al 1979) to provide ranked output. The documents that are closer to the query are ranked high (Lee 1994). The extended Boolean operators make use of the weights assigned to the terms in each document. An extended Boolean operator evaluates its arguments to a number in the range zero to one corresponding to the estimated degree to which the given logical expression matches the given document whereas a classical Boolean operator evaluates its arguments to return a value of either true or false. Boolean information retrieval has been combined with content-based navigation using concept lattices, where shared terms from previously attained documents are used to refine and expand the query (Carpineto and Romano 1998). Based on extended Boolean models, the p form and fuzzy logic approach have been developed.

In the p-form (Fox 1983) approach, a query and document term use weights for ranking the documents. Weights have been computed by using term frequency statistics with the proper normalization procedures. For an OR query, these normalized weights can be used to rank the documents in the order of decreasing distance from the point and for an AND query, the documents can be ranked in the order of increasing distance from the point. The Boolean operators have a coefficient p associated with them to indicate the degree of
strictness of the operator. A distance-based measure is used by p-form and the
degree of exponentiation is determined by the coefficient p. The advantage of
the p-form is that the evaluation procedures of all the terms are in the same way
and also the evaluation of query/document terms matching is closest to the
human’s intuition. For p-values greater than one, the exponentiation is an
expensive computation. However, expertise is needed for formulation of
extended Boolean queries.

In the Fuzzy Set Theory, traditional binary membership of an element
in a set is replaced by a varying degree of membership. The weight of the index
reflects the degree of membership of the document in the fuzzy set associated
with the term. The maximum and minimum of the degree of membership
correspond to the union and intersection of two fuzzy sets. In the model
developed by Fox and Sharat (1986), the Boolean operators are softened by
considering the query-document similarity to be a linear combination of the min
and max weights of the documents. The fuzzy operators replace Boolean
operators (Lee et al 1993). Kwon et al (1994) proposed weighted query
expansion using a thesaurus. A model based on the fuzzy set theory allows the
interpretation of a user query with a linguistic descriptor for each term
(Bordogna et al 1994, Kraft et al 1994). To provide or simplify query
formulation and enable control over relevant ranked output and address
uncertainty in query representation, statistical models have been proposed.

2.3 STATISTICAL MODELS

The vector space and probabilistic models are the two major
elements of the statistical retrieval approach. Both models use the terms with
weights but the vector space model represents the documents and queries as
vectors in a multidimensional space and compares their weights using the cosine similarity measure (Salton 1989) whereas the probabilistic model uses the probability principle to rank the documents based on the probability of their relevance to the query. Both produce their output as a list of documents ranked by their estimated relevance. Even though the extended Boolean model provides relevance ranking, expertise in query formulation is required. Hence, the vector space model, which provides relevance ranking but uses, a simple set of terms with or without weights, was proposed.

2.3.1 Vector Space Model

The vector space model represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the document (Salton 1983). The creation of an index involves lexical scanning to identify the significant terms, where morphological analysis reduces different word forms to common "stems", and the frequency of occurrence of those stems is computed. Each document is represented as a set of terms. The union of a set of terms of an entire collection forms a document space. Each term is assigned with weights, which indicate the importance of the term in that document. The term space is assigned to a given collection. In document space, each distinct term represents a dimension whereas in term space each document is a dimension. The document and term space are combined and represented by document by term matrix. Each row represents a document whereas each column represents a term. The element at row i, column j is the weight of term j in document i. The most commonly used weighting scheme is the term frequency * inverse document frequency weighted scheme, i.e. tf*idf. The term frequency is the frequency of occurrence
of the given term in the document whereas idf characterize the given term within the collection. The idf is defined as \( \ln(N/n) \) where \( N \) denotes number of documents in the collection and \( n \) denotes the number of documents that contain the given term. IDF indicates that terms, which occur in all the documents, are not likely to be useful for indexing purposes. Due to the fact each document is of different lengths, the weights of the term need to be normalized. Hence researchers have provided various normalizing schemes.

Lee (1995) proposed two types of normalizations namely augmented normalized term frequency, which divides the term frequency by 'maximum term frequency' and Cosine normalization, which is done by vector length. Both normalizations can be used separately or together. The first normalization uses the formula \( 0.5 + (0.5 \times (tf/tf_{\text{max}})) \). The drawback of this normalization is that it is fully dependent upon the maximum term frequency. Another alternative is logarithmic term frequency i.e. \( \log(rf) + 1 \). In this second method, the Euclidean length of the vector divides the component of the document vector, which is the square root of the squares of all its components. This normalization is called cosine normalization because the normalized vector has unit length and its projection on any axis in document space is the cosine of the angle between the vector and the given axis. In this technique, the normalization factors take all the vector components rather than a single term with high frequency, and then the term with a disproportionately high frequency is diluted by the weights of all the other terms. In this way, cosine normalization is taken for normalizing the weights of the term because it reduces the problem of vector component weights for a given document being distorted by a single term with high frequency.
Term weighting has been explained by controlling the exhaustivity and specificity of the search, where the exhaustivity is related to recall and specificity to precision (Van Rijsbergen 1979). In 1958, a common weighting scheme was provided by Luhn for terms in a document, which uses the frequency occurrence of the terms for weighting purposes. The weight of a term is calculated as the product of a term frequency factor, a document frequency factor, and a document length normalization factor. Using one of the weighting schemes, the document and query vectors are calculated. If both vectors have been cosine normalized, then the product of query vectors and the documents vector represent the cosine of the angle between the two vectors. This kind of similarity is called cosine similarity (Salton 1989). The zero cosine value denotes the maximum similarity whereas one denotes the minimum. The problem with cosine similarity as stated by Salton and Lee is that it tends to produce low similarity values for longer documents. Lee’s solution is to merge the result of a retrieval using cosine similarity with the result of retrieval using term frequency normalization. This model’s major limitation is that the terms are independent, orthogonal dimensions of the document space. Basically this is a term–document relationship, not a term–term relationship. The major drawback of the vector space model is that it has a large number of dimensions if the number of terms that occur in a collection is large.

A new vector space model called Latent Semantic Indexing (LSI) (Deerwester et al 1990) was developed. It tried to capture the term–term statistical relationships. In LSI, a conceptual index called LSI ‘factor’ or ‘feature’ is used in order to have lower dimensional document space called k-space in which each dimension is a derived concept. This replaces the document space in which each dimension is an actual term occurring in the
collection. The LSI factors are uncorrelated in a way that terms are not. Hence, LSI factors (Hull 1994) capture information rich term-term relationships that ordinary term based document space does not. The term independent properties of vector space model is satisfied by this model by providing the semantic term dependencies using a purely statistical and automatic method without syntactic or semantic natural language analysis and without manual human intervention. Using the matrix decomposition method called Singular Value Decomposition (SVD) provides this. The projection of a set of documents into k space is optimal in the sense that the project is "guaranteed to have, among all possible projections to a k dimensional space, the lowest possible least square distance to the original documents. In this sense, LSI finds an optimal solution to the problem of dimensionality reduction" (Schutze and Silverstein 1997).

Hull points out some drawbacks of LSI. While a reduced representation based on a small number of orthogonal variables might appear to cut storage costs substantially, the LSI values are real numbers while the original term frequencies (weights) are integers, adding to the storage costs. Another disadvantage is that the LSI is also computationally expensive for large collections. However it needs only to be constructed once for the entire collection performance at retrieval time is not affected (Hull 1994). In addition, most conventional vector space approaches use some form of query expansion to modify and expand the user's original query. LSI reduces the number of terms in the query, whereas conventional query expansion increases the number of terms in the query. LSI was compared with multidimensional scaling and it was shown that LSI preserves the document space optimally when using the inner product similarity function (Bartell et al 1992). The dimension of the transformed space in LSI is reduced by the selection of the highest singular
values. Further, the Bayesian regression model was used and it was shown that by removing the small singular values, dubious information is being removed and also specification errors are reduced (Story 1996). However the above models do not tackle the uncertainty aspect of queries and probability models were proposed to handle this issue.

2.3.2 Probabilistic Model

The probabilistic retrieval model is based on the probability Ranking Principle, which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query with the evidence available (Belkin and Croft 1992). This need arises because there is uncertainty in the representation of the information need and the documents. The retrieval effectiveness is near optimal to the evidence used (Cooper et al 1992). The actual estimates for each document of probability of relevance provided to the user are useful rather than with the ranking of documents by probability of relevance (Turtle and Croft 1991).

In a probabilistic method used in IR where query and document are represented by sets of terms, P(D/R) is calculated as a function of the probability of occurrence of these terms in relevant versus not relevant documents. There can be a variety of sources of evidence that are used by the probabilistic retrieval methods, and the most common one is the statistical distribution of the terms in both the relevant and non-relevant documents. The conditional probability may be computed based on any clues available about the document i.e. as indexed terms, concepts, synonyms etc. Probabilistic models make some statistics based simplifying assumptions. The probabilistic
formulae used to calculate $P(D/R)$ depend on the specific model used and also on the assumption of the distribution of terms in the set of relevant and non-relevant documents. Binary Independence (Cooper 1991) or Linked independence (Cooper et al 1992) is the most widely used technique. In this method, the clues to relevance are assumed to be independent of each other in both the set of relevant documents and the set of non-relevant documents (Cooper et al 1992 and Cooper 1995).

Bayesian probabilistic models relate the probability (prior probability of relevance) that a document selected at random will be relevant to the probability (posterior probability of relevance) that an observed document is relevant, given the observed features of the document. Bayesian probability can be applied to IR as a process of redistributing probabilities of relevance from the prior probability distribution over all the documents in a given collection to a posterior distribution over the set of retrieved documents. Crestani and van Rijsbergen (1995) call this as probability kinematics and view this as a flow of probabilities among the terms serving as descriptors of a document collection. The simplest of the probabilistic models is based on the presence or absence of independently distributed terms in relevant and non-relevant documents (Shaw 1995) which is the distribution of any given term over the collection of documents is assumed to be independent of the distribution of any other term. Using Bayes rule of inference and binary independence assumption, a function can be derived for ranking documents by probability of relevance (Van Rijsbergen 1979).

Turtle and Croft proposed a Bayesian inference network in 1991, which uses multiple sources of evidence to compute conditional probability.
Inference networks can be used to simulate both probabilistic and Boolean queries and can be used to combine results from multiple queries. They took multiple sources of evidence regarding the relevance of a document to a user query. Croft et al (1991) developed a model, which computes $P(\text{Relevance/Document, query})$, the probability that a user decides a document is relevant given a particular document and query. The inference net model computes $P(I/Document)$, the probability that a user's information need is satisfied given a particular document. The result is a ranked list of retrieved documents, with more traditional retrieval methods.

The statistical approaches provide users with a relevance ranking of the retrieved documents so that they enable users to control the output by setting a relevance threshold or by specifying a certain number of documents to display. Queries can be easier to formulate because users do not have to learn a query language; they can use natural language and the uncertainty inherent in the choice of query concepts can be represented. However, statistical approaches have limited expressive power. Statistical approaches lack the structure to express important linguistic features such as phrases. Proximity constraints are also difficult to express, a feature that is of great use for experienced searchers. A ranked linear list provides users with a limited view of the information space and does not directly suggest the way to modify a query if the need arises (Hearst 1995). The queries have to contain a large number of words to improve the retrieval performance. As in the case of the Boolean approach, the query terms given are the same appropriate terms as in the relevant documents. In order to over the problem where actual words in the query need to be present in the document for it to be selected rather than selecting documents based on content there is a need to use linguistic approaches.
2.4 LINGUISTIC MODELS

IR techniques are primarily aimed at detecting relevance, with little regard for linguistic phenomena. An IR technique is judged effective if it can differentiate one piece of text from another, namely a relevant document from an irrelevant one. This has been done fairly successfully using quantitative methods based on word and/or character counts, especially when relatively coarse-grained distinctions in content were sufficient. Traditional IR technology for the most part is not linguistically motivated. The rapid growth of information in web, together with developments in natural language processing (NLP) technology, has prompted those engaged with NLP to suggest that it could be usefully applied to information retrieval primarily for richer and more effective document representations. Linguistic approaches to IR essentially involve techniques based on knowledge derived from syntax and/or semantic processing of the document, or external knowledge about the world or knowledge about the domain pertaining to the document. The emphasis is for Linguistic processing to aid statistical or probabilistic methods rather than replace them. From the IR perspective, Linguistic processing is classified according to the level of linguistic unit processed as morphological, lexical, syntactic, semantic and pragmatic (Liddy 1998). A minimal extent of linguistic processing especially at the morphological and lexical levels has always been used in traditional IR techniques. Morphological stemming techniques to reduce variants of a word to a common root form, which can be viewed as a type of term normalization has long been, used in traditional IR methods. Similarly lexical methods including the construction of a list of low semantic content stop words to eliminate them from further processing and the generation and use of thesaurus for query expansion, has been used to support traditional IR methods.
It is important to emphasize that almost always NLP techniques are used in coordination with traditional Boolean, vector and statistical techniques for document representation and categorization. At the first level of linguistic processing is the processing of the text at the word level. Included here are morphological processing, parts of speech tagging and lexical processing. Part of speech taggers predict syntactic category of words using both morphological and lexical information in a text with high level of accuracy. More sophisticated word level linguistic processing is the resolution of semantic ambiguity generally known as word sense disambiguation, which has proved to be a challenging task. Syntactic based approaches break away from word based approaches and tend to group phrases based on syntactic constraints. At a higher level in the NLP analysis is the semantic based approach where the meaning of sentences or text is represented after local context analysis. The incorporation of external knowledge either in the form of conceptual knowledge or domain knowledge for analytical purpose also helps in relating documents to queries.

The best statistical or probabilistic methods will mix some relevant documents and retrieve some junk. Hence, the syntax and semantic information are processed and added as linguistic information in indexing. The linguistic information can be attached with word or phrase or words can be associated with relations.

2.4.1 Word Based Methods

The linguistic processing are morphological processing, lexical information in the form of part-of-speech tags and proper noun identification,
and sense identification. Linguistically motivated word based approaches can be classified broadly as morphological based, lexical based and word sense disambiguation methods. Words are normalized across morphological variants using a lexicon-based stemmer. The morphological level is concerned with the analysis of the variant forms of a given word in terms of its components. The lexical level is concerned with the analysis of structure and meaning at a purely word level. Tagging words with parts of speech and proper noun identification are forms of IR lexical processing. Proper names are identified for indexing, including people's names and titles, location names, organization names etc. Word sense disambiguation is essential in effective IR processing to find the correct sense of the indexed words. The sense level can be a single sense or multiple senses. Multiple senses are required in IR because incorrect disambiguation leads to poor performance (Sanderson 1990 and Sanderson, 1994). The sense can be processed with lexical, conceptual and contextual information.

2.4.1.1 Morphological and lexical based approaches

Morphological level analysis uses a stemming algorithm to reduce a word variant to its root form. Stemming is generally concerned with breaking a derived word into its prefix, root-form and suffixes. Words are reduced across morphological variants by removing prefix/suffix or by mapping a root with a root form lexicon. Porter (1980) developed a stemming algorithm, which makes conflation errors in its removal of suffixes. Krovetz (1993) designed a different stemming algorithm called KSTEM, which tried to avoid erroneous stemming, by using evidence from a dictionary, the ubiquitous LDOCE. KSTEM removed suffixes from a word variant piece by piece and after each removal, looked-up
the reduced word form in the dictionary. This stemming algorithm was designed under the assumption that if the word was in LDOCE, then the variant had a different meaning from its root form and should not be stemmed further. By considering KSTEM’s advantage of producing fully formed words as opposed to the truncated word forms from Porter stemming algorithm. Krovetz concluded that his KSTEM system was a better stemmer. Later, Xu and Croft (1996) built a stemmer that used Porter’s suffix removal rules to reduce a variant to its root, but determined the correctness of the reduction by checking with a given corpus. Checking involved computing over the entire corpus, and Co-occurrence statistics between a variant and its root. Xu and Croft’s stemmer was found to perform better when compared with Porter and KSTEM in terms of retrieval effectiveness.

Another form of lexical processing is proper noun identification. Recognizing and classifying proper nouns involve identifying which strings in a text name individuals and which classes these individuals fall into. Typical name classes include organizations, persons, locations, dates and monetary amounts. However, further classes can include book and movie titles, product names, restaurant and hotel names, ship names, etc. Proper names are recognized and classified based on partial name fragments, which are then combined to produce complete names. The system was developed originally as a part of an information extraction system and was enhanced to handle name recognition evaluation at MUC6. The recall and precision of this are both around 85% (Cowie 1996).

The precision in IR systems should increase if multiword names are treated as unitary terms and if variant forms can be linked. MUC-6 results have
shown that IE systems rely heavily on Proper Name recognition and classification. The MUC-6 and MUC-7 Named Entity (NE) Task and Multilingual Entity Task (MET) alone have stimulated the creation of more than thirty name recognition systems for a variety of languages (English, Spanish, Chinese, Japanese and Thai). The approaches adopted in these systems range from the more or less purely statistical (e.g. the use of Hidden Markov or Maximum Entropy models - BBN) to the more or less purely rule based, as well as a number of hybrid approaches, which mix statistical with rule-based techniques. Once effective approaches to identify and search proper names have been developed, there should be a heavy demand, especially for business areas, such as newspaper database where a large proposition of queries contain personal, company, product, or other names.

2.4.1.2 Sense disambiguation approaches

Besides morphological and lexical approaches, another approach, which uses syntactic/semantic techniques to improve IR performance, is Word Sense Disambiguation. The main problem with the traditional Boolean Word-based approach to Information Retrieval (IR) is that it usually returns too many results or wrong results. The usage of word senses in the process of document indexing is a much-debated field for discussions. The basic idea is to index word meanings, rather than words taken as lexical strings. A highly accurate Word Sense Disambiguation algorithm (WSD) is needed in order to obtain an increase in the performance of IR systems. Words are the basic units of reference for most of the traditional IR methods. The categorization of documents is largely dependent on the presence or absence of words. Here the assumption is that words, which are used, have the same lexical category and
posses the same sense in all documents. However, words are often associated with multiple senses i.e., display polysemy. The meaning of a word in a particular usage can only be determined by examining its context. Using the WSD process, the system avoids matching one sense of a word in a query with a completely different sense of the word in the document. Strzalkowski and Carballo (1994) have found evidence to show that linguistic approaches to disambiguate word senses can help in information retrieval.

In the early days, disambiguation was done based on manually generated rules, which were primarily used simply for Word sense disambiguation with no reference to its application for IR. Weiss (1973) manually constructed a set of rules to disambiguate five words. These rules were of two types, general context rules and template rules. To create these rules, Weiss examined 20 occurrences of an ambiguous word and then tested these manually created rules on a further 30 occurrences. These tests were performed for five ambiguous words. The accuracy of the resulting disambiguation was of the order of 90%. A large disambiguator was built by Kelly and Stone (1975) who manually created a set of rules for 6000 words. Using word experts, Small and Rieger (1982) proposed a disambiguator. Their idea was to build an expert for tackling ambiguous words. Both Kelly and Small finally came to the conclusion that building this kind of disambiguators was not an easy task. Since the late 1980s, the disambiguators are being used based on automatically generated rules with the evidence taken from the corpus. Lesk (1988) used the textual definitions of a dictionary to provide evidence for his disambiguator. He reported an accuracy of between 50% and 70%. Ranked retrieval was obtained using a scoring function based on the number of words co-occurring between a sense’s definition and the definitions of all context words. He found that the definition length is an important factor. Wilks et al
(1990) used LDOCE (Longman's Dictionary of Contemporary English) and proposed a technique of expanding a dictionary definition with words that commonly co-occurred with the text of that definition. He tested the word 'bank' which appeared in around 200 sentences. His system selected the correct sense 53% of the time. He pointed out that every permutation of word sense in a sentence needs to examine hundreds of thousands of sense combinations. Hence he suggested that the technique of simulated annealing can help handle this task because simulated annealing is a technique that can be applied to problems of combinatorial explosion. This suggestion was taken by Cowie (1992) who proposed a disambiguator, which reported 47% of accuracy, by taking 67 sentences using the senses in LDOCE. The problem with comparison of efficiency evaluation of disambiguation tasks was the absence of a standard baseline. Finally it was agreed that a baseline was needed to compare the accuracy of disambiguation. One of the baselines proposed was the measuring of disambiguation accuracy by randomly selecting senses. Ng and Lee (1996) manually disambiguated 192000 occurrences of 191 words using WordNet as a lexical resource. They took a tagged corpus with part of speech tags and then trained a statistical classifier to associate feature with occurrences of the tags. They found that the local collocation and part of speech/morphological information were the best features for disambiguation. They took 89% of corpus as a training set and the remaining as a test set. The accuracy obtained here is 63.7%. Ng and Lee selected the most common sense of a word as a baseline. Similarly Guthrie et al (1991) proposed a disambiguator based on using word co-occurrence statistics derived from another electronic resource, LDOCE MRD and obtained an accuracy of 70%. Leacock et al (1996) have investigated sense disambiguation of words in a large text corpus by statistical classification based on term co-occurrence in the context in which the given words occur. The context was taken from the current sentence and its previous
sentence since a given word is often used anaphorically. Yarowsky used Roget’s thesaurus where all the words were classified into in 1042 semantic categories. The set of clue words, one set for each category was derived from a part of speech tagged Grolier Encyclopaedia. Yarowsky (1992) reported deriving around 3000 clue words per category. The clue word sets can be compared to a context of words for disambiguation purposes. The accuracy of Yarowsky’s disambiguator was 90%.

Word Sense Disambiguation with reference to IR work was first proposed by Weiss (1973). He however reported that resolving all ambiguous words in a document collection produced only 1% improvement in retrieval effectiveness. For effective IR it was necessary to perform WSD for IR on large corpuses and this was attempted by Voorhees, Wallis and Sussna.

One approach to disambiguation for IR is the use of information from an explicit lexicon or knowledge base. The lexicon may be a machine-readable dictionary, thesaurus or may be handcrafted. Work has been done using existing lexical knowledge sources such as WordNet (Miller 1990), LDOCE (Longman 1988) and Roget’s International Thesaurus (Kirkpatrick 1988). Another approach taking evidence from a corpus for word sense disambiguation was proposed.

2.4.1.2.1 Lexical resources based approaches

Lexical resources have been extensively used in Word Sense Disambiguation. WordNet is a lexical resource proposed by Miller (1990, 1995) that has been used by a number of researchers for WSD with particular relevance to IR. Voorhees (1993) proposed a sense disambiguator based on
WordNet. Each of WordNet’s 90,000 words and phrases was assigned to one or more synsets. A synset is a set of words that are synonyms of one another. The words of a synset provided in WordNet are synonymous with the descriptions of the synset. All synsets are linked together to form a largely hierarchical semantic network based on the hypernymy and hyponymy word relations, with some additional relations of meronymy, holonymy and antonymy.

Voorhees used the synsets of nouns and the hierarchical hypernym and hyponym relations between them. Her disambiguator took the amount of co-occurrences between the word’s context and words in the synsets. Voorhees (1998 and 1999) tried to resolve word sense ambiguity in the collection of documents, as well as in the query, and then compared the results obtained with the performance of a standard run. Even though she used different weighting schemes, the overall results have shown degradation in IR effectiveness when word meanings are used for indexing. Still, as she pointed out, the precision of the WSD techniques has a dramatic influence on these results. She stated that a better WSD could lead to an increase in IR performances. Voorhees (1994) manually expanded 50 queries over a TREC-1 collection. Queries were expanded using synonymy and other semantic relations from WordNet 1.3. Voorhees found that the expansion was useful with short, incomplete queries, and rather useless for complete topic statements where other expansion techniques worked better. For short queries, there remained the problem of selecting the expansions automatically; doing it badly could degrade retrieval performance rather than enhance it.

Wallis’s (1993) disambiguator was based on Wilks’s work (1997) but was used for the purpose of IR. In addition to considering the co-occurrence of
words in context, he also considered the addition of co-occurrences of synonyms of the words since they convey the same sense, thus enabling the expansion of word set with respect to senses. He measured the retrieval effectiveness with the CACM collection and found that there is a drop in his senses based representation. Sussna (1997) expanded on Voorhees’s WordNet based work using more of the thesaurus’s semantic relations. The disambiguator resolved these occurrences with an accuracy of 56%. It was reported that a context of 41 words produced the best disambiguation accuracy.

Richardson and Smeaton (1995) attempted a similar approach of replacing words in a document collection with a representation derived from the WordNet Semantic network. They did Query expansion with WordNet, which was shown to be potentially relevant to enhance recall, as it permits matching relevant documents that could not contain any of the query terms. This automatic disambiguation and measures of semantic relatedness between query/document concepts resulted in a drop of effectiveness. Unfortunately, the effects of WSD errors could not be discerned from the accuracy of the retrieval strategy.

Smeaton and Quigley (1996) used the same approach of using an expanded sense based representation of a collection, but found significant improvements in effectiveness. Smeaton and Quigley’s work (1996) showed that using senses to help in term expansion is potentially a good strategy. His retrieval on very short documents is reasonably improved based on WordNet 1.4. These results are in agreement with Voorhees (1994), but the question remains whether conceptual distance matching would scale up to
longer documents and queries. In addition, the implementation in Smeaton and Quigley (1996) considered only nouns, while WordNet offers the possibility of using all open-class words (nouns, verbs, adjectives and adverbs).

Gonzalo et al (1998) expanded on Voorhees WordNet based work using additional semantic as in Sussna’s work. However Gonzalo proposed the use of weights to prioritize Semantic relations and calculate term distance. In this work, each synset was given a score calculated as the sum of semantic distances between the context words and that synset. The score was used to rank the synsets, with the top one being chosen as the correct sense of the word. Because synonyms of a word sense were part of the same synset, Gonzalo et al believed that the representation would be richer and the accuracy was found to be 56%. Gonzalo et al performed experiments in sense based indexing where they used the SMART retrieval system and a manually disambiguated collection. It turned out that indexing by synsets could increase recall up to 29% with respect to word based indexing. Part of their implementation was the simulation of a WSD algorithm with error rates of 5% 10%, 20%, 30% and 60%. It was found that error rates of up to 10% do not substantially affect precision and a system with WSD errors below 30% still performs better than a standard run. The results of their experiments were encouraging, and proved that a WSD algorithm with better accuracy can significantly help IR systems. Clearly the grouping of synonyms, which the synset representation provides, had a positive impact on the retrieval situation.

Further, Schutze and Pedersen (1995) performed experiments, which have shown that semantics can actually help retrieval performance. They reported an increase in precision of up to 7% when sense based indexing
is used alone, and up to 14% for a combined word based and sense based indexing. One of the largest studies regarding the applicability of word semantics to IR was reported by Krovetz & Croft (1993, 1995 and 1997). When talking about word ambiguity, they collapse both the morphology and semantic aspects of ambiguity, and refer to them as polysemy and homonymy. They showed that word senses should be used in addition to word based indexing, rather than indexing on word senses alone, basically because of the uncertainty involved in sense disambiguation. They had extensively studied the effect of lexical ambiguity over IR and the experiments described provide a clear indication that word meanings can improve the performance of a retrieval system. Kroverz and Croft examined the improvement in retrieval effectiveness caused by removing retrieved documents, which contained sense mismatches. On the Time collection, a 4% increase in average precision was found, but on the CACM collection the increase was 33% although Kroverz and Croft put much of this large improvement down to better stemming while the sense matching was being performed.

Mihalcea and Moldovan (2000) proposed a WSD method for open domains and observed that 55% of the nouns and verbs were disambiguated with a precision of 92.22%. He used a hybrid indexing approach that combined Word-based and Sense-based indexing, which called Semantic is indexing. The relative gain of the combined word based and synset-based indexing with respect to the basic word-based indexing was 16% increase in recall and 4% increase in precision. The additional use of hypernym synsets resulted in 28% increase in recall, with a 9% decrease in precision.
Krovertz and Croft, Sanderson and Gonzalo carried out evaluation and analysis of the effect of an IR in disambiguation. They provide some clues to the reasons for the lack of success. Krovertz found that the collocation effect and the skewed frequency distribution of word senses affect their performance. He also concluded that the disambiguation was ineffective where the effects of collocation were less prevalent and the query words have senses with uniform distributions or were used in a minority sense. Sanderson (1994, 1997) used pseudo-words to test the utility of disambiguation in IR. Different levels of ambiguity were introduced in the set of documents prior to indexing. The conclusion drawn was that WSD has little impact on IR performances, to the point that only a WSD algorithm with over 90% precision could help IR systems.

Sanderson’s analyses have echoed some of Kroverz and Croft’s work and have also investigated the impact of erroneous disambiguation on IR effectiveness. Sanderson also used pseudo words to study the impact of automatic disambiguation on retrieval effectiveness concentrating particularly on the impact of disambiguation errors. By resolving the ambiguity in the pseudo words and occasionally making a mistake, he found that a 20-30% error rate could cause effectiveness to be as bad as or even worse than when ambiguity was left unresolved. Sanderson concluded that a disambiguator was only of use in a retrieval context if queries were short or if disambiguation was performed at a high level of accuracy. The reasons for the results obtained by Sanderson have been discussed in Schutze and Pedersen’s work (1995). They argue that the usage of pseudo-words does not always provide an accurate measure of the effect of WSD over IR performance. It is shown that in the case of pseudo-words, high-frequency word types have the majority of senses of a
pseudo-word, i.e the word ambiguity is not realistically modelled. The analysis of the results concentrated on the length of queries showing that the effectiveness of retrievals based on a query of one or two words was greatly affected by the introduction of ambiguity but much less so for longer queries, confirming the collocation effect noted by Krovetz and Croft. Sanderson concluded in his summary that the three factors affecting disambiguation in IR are, the skewed distribution of the senses of many words along with word collocation effects, the better accuracy of a disambiguator and fine grained senses needed in a dictionary or thesaurus. Sanderson also used a set of representations composed of all the senses each with their assigned score. He found that retrieval on documents represented by a full sense ranking was better than retrieval on the representation where a single top ranked sense was used. In contrast to the pre-defined set of word sense definitions, another approach to disambiguation in IR is based on senses derived from corpora.

2.4.1.2.2 Corpus based approaches

Corpus based approaches disambiguate words using information which is gained by training on some corpus, rather than taking it directly from an explicit knowledge source. This training can be carried out on either a raw or disambiguated corpus. A disambiguated corpus is one where the semantics of each polysemous lexical item is marked with correct sense, whereas, a raw corpus is one without such marking. In general, machine-learning algorithm of some kind is applied to certain features extracted from the corpus and used to form a representation of each of the senses. This representation can then be applied to new instances in order to disambiguate them. A common feature set used by Kilgariff (1991) was to take all the words in a window of words around
the ambiguous words, treating the context as an unordered bag of words. Another approach used Hidden Markov Models, which have proved very successful in part-of-speech tagging. Realizing that semantic tagging is a much more difficult problem than Part-of-speech tagging, it was decided to perform an experiment to see how well words can be semantically disambiguated using techniques that have proved to be effective in part-of-speech tagging. The lack of tagged resources has led to the explosion of the use of unannotated, raw, corpora to perform unsupervised disambiguation. Dynamic matching techniques examine all instances of a given term in a corpus and compare the contexts in which they occur for common words and syntactic patterns. A similarity matrix thus formed is subjected to cluster analysis to determine groups of semantically related instances of terms.

Apart from the presence of a predefined set of word sense definitions, another approach to disambiguation in IR was using senses derived from the corpus. Using disambiguation in IR, Zernik (1991) decided to cluster word occurrences based on their contexts taken from the document collection. Once generated, Zernik would attempt to associate the clusters with some dictionary senses. Using twenty disambiguated words within a corpus, Zernik performed retrieval and examined the change in retrieval accuracy and stated that accuracy was improved by 50%. Zernik (1991) concluded that sense definitions in a dictionary are based on grammatical rather than semantic criteria.

Schutze and Pedersen (1995) improved retrieval effectiveness further, using a corpus as evidence. For each word to be disambiguated in the corpus, the context of every occurrence of that word with in the corpus was examined and similar contexts were clustered on context word alone According
to his disambiguator, each one of the similar contexts constituted an individual sense of the word, making them quite different from senses defined in a dictionary. They showed 14% improvement in retrieval effectiveness.

2.4.2 Group Based Methods

Phrases are typically identified in IR so that they can be used as terms that are the terms of a document extended to more than a single word. There are a number of ways to obtain “phrases” from the text. These include generating simple collocations, statically validated N-grams, Part-of-speech tagged sequences, syntactic structures, even semantic concepts. The syntactic level is the level at which the syntactic structure of a sentence is determined, in terms of the Part-of-Speech of the individual words. Syntactic level processing has been used to identify phrases by statistical co-occurrence and proximity rather than Natural Language Processing techniques. Noun phrase recognition is based on first using a Part-of-Speech tagger.

Traditional methods identify phrases by statistical co-occurrence. Syntactic analysis can identify phrases even when the terms of which they are composed are not adjacent, or do not co-occur with greater than chance frequency. A combination of syntactic and statistical methods is more effective (Lewis 1992, Lewis et al 1996). However, they suggest, “Weighting for phrases may differ from weighting for single word terms to allow for their lower frequency and different distribution characteristics”. Lewis et al also identified that the degree of NLP and statistical processing applied to extraction of phrases and other compound terms might be considerably greater for user queries than for documents.
In Text Retrieval Conference-2 (TREC-2), Strzalkowski and Carballo (1994) extract phrases syntactically from a large collection, but they apply a variety of statistical techniques to these phrases before formulating queries. The extracted phrases are statistically analyzed in syntactic contexts in order to discover a variety of similarity links between smaller sub phrases and words occurring in them. A further filtering process maps these similarity links onto semantic relations (generalization, specialization, synonymy etc) after which they translate a user’s request into a search query. Syntactic phrase extraction is performed and used somewhat differently by Riloff (1995). He extracts instances of phrases patterns called concept nodes, which can be used as descriptor terms in IR. He obtained 95% precision for the test collection. Prepositions, passive vs active verb form, and positive vs negative assertion, also proved significant in determining the significance of a phrase as a descriptor, at least within a specific domain.

Dillon and Gray (1983) applied a shallow processing strategy to obtain complex terms. They used tagging, reducing tag ambiguity by tag-string constraints and applied a set of rules defining acceptable tag sequences in order to select phrases. Sparak Jones and Tait (1984) used NLP to determine sentence propositional structures from which complex terms could be extracted. Smeaton and Van Rijsbergen (1988) identified nominal phrases, with co-location of phrase elements as the matching condition on documents. Fagan’s work (1987) was to evaluate non-syntactic, i.e., joined terms under alternative settings for system parameters involved, eg, degree of proximity, frequency properties of term members etc.
The recognition of a specified proper noun is essential in IR, particularly for a document to be relevant to a specified topic. Recognition, extraction and matching of a proper noun is essential in IR particularly for a document to be relevant to a specific topic. Many proper nouns consist of more than one noun and also include a preposition. Many proper nouns can be specified in multiple forms. Many proper nouns are group nouns, which may result in references either to the group as a whole or to the individual entities making up the group. Common nouns and noun phrases may also group individual entities that have proper noun names.

Paik et al (1996) developed a sophisticated series of procedures for proper noun recognition and matching in their DR-LINK (Document Retrieval through LINguistic Knowledge) and KNOW-IT (KNOWledge base Information Tools) IR engines. DR-LINK also expanded group proper and common nouns so that a topic request could match a document on either the group name or its constituents. Various shallow text processing techniques such as part-of-speech tagging, phrase boundary detection and word co-occurrence metrics are used to identify relatively stable groups of words like joint venture. 'Head + Modifier' pairs are identified in order to normalize across syntactic variants such as weapon proliferation, proliferation of weapons, proliferate weapons etc and reduce to a common 'concept' e.g weapon+Proliferate.

Noun phrases identified as groups can then be used as descriptor terms in IR to improve performance. The knowledge about the document in terms of words, phrases, concepts and relations can be used as indexes to represent the documents for retrieval. The semantic content can be added with the terms for richer representation.
2.4.3 Semantic Based Methods

A user really wants to retrieve documents about certain concepts irrespective of the surface words that appear in a text to realize these concept. BADGER is a text analysis system, which uses linguistic context to identify concepts in a text (Riloff and Jones 1999, Riloff 1996). Single words taken out of context may not relate to the same concept as the phrase to which that word belongs. The system aims to find linguistic features that reliably identify the relevant concepts, representing the conceptual content of a phrase as a case frame or concept node (CN). CRYSTAL automatically induces a dictionary of CNs from a training corpus. The CN definitions describe the local context in which relevant information may be found, specifying a set of syntactic and semantic constraints. When the constraints are satisfied for a portion of a text, a CN is instantiated (Soderland 1995).

Woods (1997) uses semantic relationships among concepts to improve IR. NLP and Knowledge representation techniques are used to deal with differences in terminology between query and target. He developed a prototype system for indexing and organizing information in structured conceptual taxonomies. Comparing the performance of the system that uses conceptual indexing, with the performances obtained using classical retrieval techniques, resulted in an increased performance and recall. He also defines a new measure, called success rate which indicates if a question has an answer in the top ten documents returned by a retrieval system. The success rate obtained in the case of conceptual indexing was 60% with respect to a maximum of 45% obtained using other retrieval systems. This is a significant improvement and shows that semantics can have a strong impact on the effectiveness of IR.
systems. The FEERRET systems (Mauldin 1991) is another example of how concept identification can improve Information Retrieval Systems

Chakravarthy and Haase (1995) proposed a technique where the semantic contents were associated with structured descriptors in achieves for representing the documents and queries. The Semantic level is the level at which one tries to interpret meaning at the level of clauses and sentences rather than just individual words. The words associated with relations are extracted from the text. The relations formed can be used as one more dimension in indexing.

The effectiveness of Word/Sense/ Phrase-based retrieval systems is mostly poor because the system retrieves only the documents that contain those terms that occur also in the query. The word context is one way of improving the information retrieval performance. The context can be local / micro context or topical context. The micro context is the context generally surrounding a word occurrence in a text, essentially a window around the word under consideration limited to a single sentence (Holub and Bohmova 2000).

There is a need for developing an IR, which can handle incomplete and inconsistent queries, and also inferring more information. It has been suggested (Huibers 1996, Van Rijibergen 1986, 1989) that such formalisms can be both appropriately and powerfully defined within a logic (Barwise 1993).

2.5 LOGICAL MODELS

Information Retrieval needs to include formalisms that can deal with uncertainty, and as a tool to capture imprecise knowledge and to reason on that
knowledge, there is probably no better formalism than logic. The most promising IR formalisms are those which combine consistently well-known approaches for assessing uncertainty such as probability theory, with paradigms that can represent uncertain knowledge and allow inferences on it. Rijsbergen's logical uncertainty principle underlies most logical IR models.

Logic has been used for many years in artificial intelligence to formalize the manipulation of information by intelligently using information to think, infer, conclude and acquire knowledge and take decisions. A primary aim of the IR system is to capture the manipulation of information, during retrieval process or query evaluation. Logic makes it possible to reason about an IR system and its properties Huiber's (1996) objective was to determine an appropriate logic and a theory of uncertainty and to develop a method to combine them in a consistent fashion. This choice has the advantage that it leads naturally to a numerical expression of relevance. Many of these logic models are based on the Logical Uncertainty Principle (Van Rijsbergen 1986). The logical model takes hypertext links or other inter document relationships into account in order to enhance retrieval effectiveness. Logic was instantiated by the formalism of conceptual graphs (Sowa 1984), which are graphs built out of concepts and their associated semantics.

Terminological Logic provides an object oriented flavoured knowledge representation. The primary syntax starts with terms, which are either individuals or relations. Concepts are defined on top of those (Meghini et al 1993, Sebastiani 1994). Documents were represented by individual constants, whereas a class of documents was represented as a concept. The fact that a particular individual was an instance of a concept was
written as an assertion. The implementation of logical system can be complex and when possible, only small document collections have been handled. However, implementations of standard test collections have been recently attempted with some positive results (Crestani and Van Rijsbergen 1995).

The semantic representation associated with each index term is validated by domain ontology. A tool was proposed for extracting ontologies from the texts (Todirascu 2001). The ontology was described using Description Logic (DL), selected due to its properties of handling incomplete and erroneous data and the associated logical inference mechanisms. This method identified the potential term candidates in the texts, using shallow natural language processing tools. Relations between terms and concepts were described by a set of rules based on shallow syntactic knowledge and on domain ontology (Todirascu 2001).

2.6 WEB INFORMATION RETRIEVAL

Many software tools are available for Web Information Retrieval, such as search engines, hierarchical directories, many other software agents and collaborative systems. In classic information retrieval, the performance of an IR system is evaluated along three lines: recall, precision and precision at 10 result pages. Though, web IR provides high reliability, it is also required to return high quality (Valuable) pages. With the available search engines, user needs are not satisfied with the different formats for inputting queries, speeds of retrieval, presentation formats of the retrieval results and quality of retrieved information (Lawrence and Giles 1998). One of the main problems with classical searches is the inability to find relevant information. Several different measures have been
proposed to qualitatively measure the performance of classical information retrieval systems (Losee 1998, Manning and Schutze 1999). Most of the searches are taking the measures of web information retrieval as precision and recall.

Caching documents to reduce access delay is extensively used on the web. Automatically identifying mirrored collections on the web can improve the performance of crawling, ranking, archiving and caching (Kobayashi and Takeda 2000). Hierarchical directories are the ontologies of the Web. They provide a more focused way of document searching. Instead of searching the entire Web, query can be restricted to a particular relevant category and in this way the quality and the relevance of the result set can be improved.

The grouping together of similar documents to improve information retrieval is known as clustering (Anick Vaithyanathan 1997; Rasmussen 1992; Sneath and Sokal 1973; Willett 1988). Anick Vaithyanathan combines the results from LSI and analysis of phrases for context based IR on Web. Zamir et al (1997) developed three clustering methods, word intersection clustering method, phrase intersection clustering, and suffix tree clustering methods for documents. Modha and Spangler (1999) developed a clustering method for hypertext documents, which uses words contained in the document, outlinks from the document and inlinks to the document. Clustering is one of several ways of organising documents to facilitate retrieval from large databases.

The Boolean, Vector space model and LSI are used by the current search engines. Statistical approaches used natural language modelling and IR
can be used in web information retrieval. These approaches are proposed and analysed in Crestani et al. (1998) and Manning and Schütze (1999). An intelligent agent proposed by Finin et al. (1998) includes the task of finding and filtering information, customizing information and automating completion of simple tasks.

The main problem with the current search engines is the large volume of documents extracted as a result of broad, general queries and the lack of output produced to specific, narrow questions (Selberg and Etzioni 1995). Baeza and Ribeironeh (1999) proposed an intelligent agent for indexing. The metadata in the context of web pages on the Internet refers to an invisible file attached to a web, which facilitates collection of information, by automatic indexers (Huang 2000).

Two main approaches have been considered by researchers in trying to improve the quality of the search on Internet or large collections of texts. The first one is to make use of multiple search engines and create a Meta search engine (Selberg and Etzioni 1995, Gravano 1997). The second approach is to use Natural Language Processing Techniques. Machine Readable Dictionaries have been used for query extension to increase the number of documents retrieved. This method has been developed for information retrieval on the Internet or in very large text collections.

Berkeley (1999) did a survey on performance evaluation for various existing search engines like google, Alta Vista, Northern light, Infoseek and Fastsearch. Google incorporates an innovative ranking algorithm in result page ranking. It provides a relatively more relevant, high quality result than others.
because of this ranking mechanism. To deal with large volume of data, Google uses asynchronous I/O to manage events, and a number of queues to move page fetches from state to state.

AltaVista has the largest data collection among all the existing search engines. Northern lighting is better serving queries on academia and business topics. Infoseek distinguishes itself with a searching within results feature, which is very powerful and as good as an excellent ranking mechanism. Fast search has the second largest data collection as well as the fastest search engine.

The NLP based system (REASON and INQUIRY) is used to improve the quality of the information retrieved using NLP based systems (Berry and Browne 1999).