CHAPTER 8

CONCLUSION AND FUTURE ENHANCEMENTS

8.1 CONCLUSION

The main objective of this thesis is to design and implement semantic oriented methodology to improve recall and precision in IR. Towards the end semantics has been utilized at various linguistic levels and at various stages of the IR process. At the word level, additional semantics has been used to improve the word sense disambiguation algorithm, resulting in more accurate sense determination of the indexed words. Incorporation of semantically related sentence level constituents as part of the index structure, allowed enriched representation of the document. A concept oriented logical level representation of the document also served as enrichment to the document representation, in addition to allowing inferencing about the contents of the document. In this way, semantics has been introduced at the word level, sentence level and document content level to enhance the representation of the document resulting in improved IR performance.

To start with, it was decided to improve the performance of word-oriented approaches to indexing in IR. Keeping this as the basis, the performance of sense based indexing has been improved by improving the accuracy of the word sense disambiguation algorithm. Additional semantics in the form of concept and context has been used at the word level to improve the
accuracy of the WSD algorithm. Three semantic WSDs viz. concept based, context based and priority based weighted WSD have been implemented. In concept based WSD, in addition to lexically related words, conceptually related words are also considered to improve the performance of IR. These conceptually related words are obtained using glosses of WordNet. The correct sense of the words obtained using the above-described concept based WSD, has been used for semantic chaining, where a set of semantic chains represents the documents. The document clustering process, which normally uses lexical chaining (Green 1999, Hirst and Morris 1991), has been modified in this thesis using semantic chaining (Manjula and Geetha 2003a) where only the synsets of the correct sense of the word are used for chaining. The set of semantic chains has been used for clustering the documents. Pre-clustering of documents before retrieval improves the performance of IR.

In the second approach to word sense disambiguation, contextual information of words which is also important in finding the correct sense of a word is considered. In this context based WSD, the context window words corresponding to the indexed words are taken and the commonly occurring semantically related words of the context window words are retained. A graph-based approach is used to find the correct sense of the word.

The degradation in performance due to incorrect disambiguation has motivated the design of a third WSD algorithm. With this idea in mind, multiple probable senses of a word are maintained with priority-based weights. The priority based weighted WSD algorithm assigns weights to the probable senses of the word using 3 levels of co-occurrence, namely co-occurrence at the synonym synset level, co-occurrence at the Hyponym synset level and
The accurate senses of the indexed words determined by the concept or context based WSD, or multiple senses determined by priority based weighted WSD are utilized in the indexing phase of IR to improve precision and recall of the IR system.

Sentence level semantics is obtained by considering thematic roles, extracted from the documents. A three dimensional indexing of the document has been implemented where the first and second dimension, word and corresponding sense are associated with the word level while the third dimension, thematic role relations is associated with the sentence level. The thematic roles extracted through appropriate heuristic rules are used to construct sentence constituents that are then used for indexing.

The document content level semantics is obtained by extracting knowledge units utilizing predefined templates or schemas. These templates enable important components of the documents to be extracted. The templates are instantiated through appropriate goal oriented extraction of information. The knowledge so extracted is represented as predicates when predicate logic is used as representation mechanism and as terminological concepts and instances when description logic is used as representation mechanism. The predicates of the predicate logic mechanism are represented using a database and then appropriate integrity constraints and semantic rules available in the knowledge base are utilized to optimally query the database. In the other representation, formalism description logic has been chosen to represent the schema of the document. The knowledge of the document is saved in the terminological box and assertion box of the DL representation. With this structured representation,
The template hierarchy is maintained which enables inferencing in the information retrieval process. With reasoning services available in this approach, inconsistent and incomplete queries are serviced.

The major contributions of this thesis may be summarized as follows:

➢ Improving the performance of Information retrieval systems

- Use of semantically motivated approaches to IR improves its performance in terms of precision and recall. The precision and recall of all the approaches have been compared with word based and other existing approaches and it was found that in general the performance of IR has been improved.

➢ Use of additional semantics to improve the accuracy of WSD

- The additional semantics in the form of conceptual and contextual information in WSD process improved the accuracy of WSD. Using these semantic WSDs, the meaning of the words used for indexing in IR, has been chosen more accurately. This ultimately improves the performance of IR. The performances of semantic WSDs in IR have shown better performance compared to word and existing sense-based approaches. In concept based WSD for IR, there is an increase of 26% precision compared to the word based approach and an increase of
13% precision compared to the sense based approach. In context based WSD for IR, there is an increase of 16% precision compared to the word based approach and an increase of 5% precision compared to the sense based approach. The recall values of semantic WSDs have also shown improved results.

- Use of prioritized multiple senses for improving search engine performance

- Multiple senses obtained from priority based weighted WSD have been used in the indexing phase to improve the performance of the semantic search engine. The indexed words of the documents are disambiguated by the priority based weighted disambiguation algorithm. The multiple senses of a word retained by this algorithm introduce additional semantics in the indexing phase of the search engine and thus effectively eliminate incorrect disambiguation. The performance of the search engine is measured as P@N (Precision at N). This denotes the number of relevant documents that occur in the first N documents of the retrieved results. The precision@10 has been used as a measure of performance evaluation. The performance of the semantic search engine is measured with the word based approach and compared to the four commercial search engines such as Google, Yahoo, Hotbot and Lycos and it was found that 40% and 16%
increase in precision of semantic search engine compared to word based and google and also it was shown with better results for the other search engines.

Going from word level to knowledge content level

- Thematic roles, which have been extracted from the sentences of the documents, have been added as a third dimension in semantic indexing of IR to improve its performance. The representation of document is enriched with additional information in the form of thematic roles, which in turn provides the semantic representation of sentence constituents rather than only words or senses. These relations have been used in semantic indexing and show better performance when compared to the word based approaches. There was 60% increase in the precision of relation based IR compared to the word based approach.

- The use of semantic based information extraction for extracting the content of documents resulted in an enriched representation of the document. Thus the representation of the document moved from word level representation to knowledge content level representation. The knowledge level representation also allows inferencing with the hierarchical representation and optimized semantic query processing with semantic rules and integrity constraints.
Use of semantically motivated approaches at various stages of IR

- The document and query are represented with correct senses with the semantic WSDs. The representation has also been enriched with relations to utilize sentence constituents rather than only words or senses in the indexing phase of IR. These semantically enriched representations of the document resulted in the improvement of IR performance.

- The semantics in the form of integrity constraints and semantic rules in the retrieval phases of IR allows optimized semantic query processing which in turn reduces the execution time of the query.

Due to incorporation of additional semantics, resulting in enriched and appropriate information in representation and indexing, the matching between query and documents is more accurate, which produces better retrieval performance. Thus it can be concluded that semantically oriented techniques improved the IR performance in terms of recall and precision.

8.2 FUTURE ENHANCEMENTS

Semantically motivated approaches can be adapted to all tasks of the text-mining framework to improve performance. Text mining is a multi-disciplinary field, which includes information retrieval, information extraction, text summarization, clustering and categorization. The use of a limited amount of NLP to extract salient semantic features of a document can help in all tasks
of text mining especially in IR and IE. An attempt to apply statistical approaches to direct NLP can allow for more effective processing for these tasks. Using a combination of semantic and statistical approaches, summaries can be generated in an efficient way. Similarly, semantic class based hierarchical approaches can improve the performance of text categorization and clustering systems. Semantic extraction of association rules from the documents can produce appropriate patterns in text mining.

Use of Interlingua approaches for representation of document content can allow for Multilingual Information Retrieval. Two approaches can be used in such cases. One is the use of Universal Networking Language (UNL), and the other is the use of Description logic for representation of semantic concepts of the documents. In such cases there is a need to design a new language independent indexing methodology for this purpose.