Chapter 7

Conclusion

With huge amount of information available on the World Wide Web, there is a pressing need to have information access systems that would help users in providing the relevant information in a concise, pertinent format. It is becoming very difficult to generate meaningful and timely summaries, with the increasing volume of online information.

In this thesis we proposed a novel procedure for ‘Automatic English Text Summarization’, in which it not only gave good summaries, but also is efficient in summarizing ‘technical’ content and the performance is satisfactory in ‘general’ content also.

In this thesis we explored the problems facing Indian language information access and quantified the extent of the problem. Our experiments on text summarization tasks using state of the art algorithms used for English like languages showed low accuracy when applied to Indian languages like Telugu.

So we presented our approach towards ‘Single and Multi-document Telugu Automated Text Summarization’. Our approach attempts to generate a text summary from the articles of Telugu news
papers, while avoiding the repetition of identical or similar information and presenting the information in such a way that makes sense to the reader. The experiments are done by taking the corpus, which was developed by typing the articles from Eenaadu newspaper over 100 days and converted to the font-encoded pages into ISCII standard encoding. The corpus includes 1970 articles totaling to 4.5 Million words. Of the 1970 documents in the Telugu News Articles corpus, 300 documents in 4 major categories (P-Politics, S-Sports, E-Editorial, and C-Cinema) have been used in the current set of experiments. The accuracy was found to be 91% when compared with machine generated to human generated summary. We observed that about 95% of Indian language content on the web was not summarizable due to proprietary font encodings, leading to low accuracy.

We also proposed ‘Ontology Based Automatic Text Summarization’, in which we present an approach to sentence extraction that maps sentences to nodes of a hierarchical ontology. By considering ontology attributes we are able to improve the semantic representation of a sentence’s information content. Our experimental results show that the ontology-based extraction of sentences outperforms baseline classifiers, leading to higher accuracy of summary extracts.