CHAPTER 9

CONCLUSION AND FUTURE SCOPE

This chapter sums up the outputs of the research and the achievements gained through it. Author's view of the achieved results is discussed in detail with the possible future works that can be carried out to improve the quality of summary.

9.1 Overview

The growth of digital information in a decade has led to the problem of information overload. Text analytics for such data presents many new challenges for research and development, and has also gained interest from industry. Automatic text summarization is a well-known solution to the problem of information overload. Text summaries are an essential guide to the users to form an opinion on the relevance of the document. In other words, summaries save time of internet users in their daily work.

From literature survey it was observed that most of the existing summarization systems have been built either on statistical approaches or on linguistic approaches. Statistical techniques started with shallow features such as term frequency (tf-idf) and gradually extended to positional features and domain-specific thematic features to improve the quality of summary. The statistical techniques were found to be simple and faster in implementation. They worked efficiently with larger documents also. The statistical techniques lacked in semantic analyses of the textual units and thus generated summary that lacked cohesiveness and coherence.

The linguistic techniques explore the discourse structure of the document by using semantic analyses of the text. It needs the support of Lexical database to find the relatedness (connectivity) of the textual units. This technique generates cohesive summary as compared to statistical techniques using shallow features. It has high complexity level of implementation as compared to statistical techniques and works slower for large

documents. It is not useful for domain-specific summarization as it does not use domain-specific features.

To achieve the benefits of statistical and linguistic methods a hybrid approach is used to generate a summarization system that uses semantic analysis of document along with important features of textual units in news domain and anaphora feature for resolution of correlation of sentences

To study the effect of the hybrid approach, three methods have been implemented separately and tested for the same datasets:

- Text summarization using lexical chaining

The advantage of this Linguistic method is: It is appealing because it offer perspectives for more semantically and linguistically rich treatment of text for summarization. Lexical chains help to capture all the sentences related to the central theme of the document providing the coverage of the topic and thus cohesive summary is generated. The limitation of this method is: It does not consider domain-specific features and hence cannot be used for domain-specific summarization. It works efficiently for small documents.

- Text summarization using fuzzy logic

This method uses feature extraction and fuzzy logic for decision module. Fuzzy logic handles the uncertainty and impreciseness in feature extraction. The framework can be extended to handle any number of features without any major changes. Its speed is high and can work efficiently on large documents also. The limitation of this method is the lack of semantic analysis due to which summary generated may not be cohesive. It does not handle correlation of sentences.

- Text summarization using LexicalFuzzySum: An hybrid approach that results in an efficient domain-specific text summarization using lexical chaining and fuzzy logic. It uses combination of statistical and Linguistic methods. It also includes anaphora resolution to handle correlating sentences.

9.2 Contributions of research

The research has focused on feature extraction for finding good combination of features, co-reference resolution for achieving coherence, integration of semantic analysis for achieving cohesiveness and optimization of rule-base, .

- The thesis presented an approach where sentences are modeled as a set of features. The features capture the statistical, linguistic and correlation aspects. An extensive analysis of the feature sets was carried out to understand their impact on capturing information as shown in chapter 7. The number of features was reduced to optimize the size of rule-base used as decision module. With the reduction of features, the quality of summary generated was still found to be much better than Baseline & MSWord in terms of precision recall and F-measure as shown in table 9.1. A good combination of features consisting of statistics of document like TF-ISF, sentence length, sentence position score, numerical data, sentence-similarity, Thematic-words score and linguistic feature chain-weight score is used.

- The thesis has focussed on anaphora resolution for the correlation of sentences in the document. This is a linguistic feature which picks up the preceding sentence for the summary if the next sentence selected for summary starts with anaphora. This approach considers the feature scores of the preceding sentence too instead of directly selecting it for summary. Correlating sentences are included into the summary to make the summary coherent.

- The thesis has highlighted the importance of using the semantic analysis of the document through semantic chaining using Word net. The chain weight of each sentence is considered as an important feature in feature extraction module. This helps to achieve the cohesiveness in the summary as shown in the figure 7.8. The integration of chain weight in feature extraction has not come to publish according to author's knowledge.

- The rule-base is again optimized by eliminating the rules written for the unimportant sentences due to their least impact on the generation of summary observed from results. This reduced the substantial storage space of rule base (by 20%). and the time required for comparison with all the rules in the rule base to decide the degree of importance of each sentence in the document as shown in 7.4.6.1 of chapter 7.

9.3 Conclusion

The system has developed an automatic text summarization system using natural language processing technique i.e. semantic chaining in combination with feature extraction using fuzzy logic and handling of correlation of sentences. It has used a dataset of 250 small-sized documents and 250 medium-sized documents in domain of sports and technical from BBC news corpus for testing the performance of the implemented summarizer.

This framework is automated to work with the articles of any domain with change in the list of thematic words. It works on the files stored on the secondary memory and on URLs used for fetching and summarizing textual documents on the web .The LexicalFuzzySum presents a more general framework where all features are easily pluggable into our framework, thereby providing more flexibility.

Experiments have supported our intuition and our system is compared with hard-to-beat baseline and MS Word summarizers in the area of text summarization. The genericness of our features with respect to their applicability to text summarization differentiates it from any of the existing approaches for summarization.

The system uses Java version of ROUGE n-gram as an evaluation tool for automatic summarization. It measures the quality of summary using metrics precision, recall and F-measure.. The experimental results and analysis of our approach is presented in chapter 8.
The table 9.1 shows the range of average F-measure values for small-sized documents and medium-sized documents of all the summarizers used for comparison.

The LexicalFuzzySum gives average F-measure values of small-sized documents in the range of 68% to 73% which is high as compared to all other summarizers in the table. Similarly for medium-sized documents LexicalFuzzySum gives average F-measure in the range of 55% to 61%. It concludes that LexicalFuzzySum is an efficient summarizer and gives better performance as compared to other summarizers in the table which concludes the first contribution on research.

Summarizer	Small-sized documents	Medium-sized documents
Baseline	48% to 65%	35 % to 47 %
MS Word	50% to 63%	40% to 55%
Lexical chaining approach	55% to 60%	45 % to 55%
Fuzzy logic	58% to 67 %	40 % to 48%
LexicalFuzzySum	68% to 73 %	55% to 61%

Table 9.1 Average F-measure values of LexicalFuzzySum, baseline, lexical, fuzzy & ms word

9.4 Scope for Future Research

In this section, we mention some of the possible future extensions of this research. In this thesis, we focused on summarization of news articles belonging to sports and technical domain. The techniques proposed here are adaptable across other domains.

One of the future plans may be to apply the topic-focused summarization framework to news articles or blogs and to extend the work in the machine leaning approaches. Topic-focused summaries of news articles would be lot more accurate and valuable to users. It would be more interesting to work on topic modeling and summarization in the domain of social media in future.

The rate at which the information is growing is tremendous. Hence it is very important to build a multilingual summarization system and this research could be a stepping stone towards achieving that goal provided there is availability of online lexical databases in other languages. The work presented by the thesis can also be applicable to multi document summarization by using minimal extensions.

The thesis has used evaluation metrics Precision, Recall and F-measure to measure performance gain over existing systems with ROUGE tool. In future work, new metrics can be investigated which can be used in automatic evaluation environment to measure the overall quality such as grammar, readability, prominence and relativeness.

The state of the art summarization systems are all extractive in nature, but the community is gradually progressing towards abstractive summarization. Although a complete abstractive summarization would require deeper natural language understanding and processing, a hybrid or shallow abstractive summarization can be achieved through sentence compression and textual entailment techniques. Textual entailment helps in detecting shorter versions of text that entail with same meaning as original text. With textual entailment we can produce more concise and shorter summaries.

The Implemented system in this thesis can work as framework for the research community to understand and extend the applicability of cognitive and symbolic approach in various domains of business needs.

Research in summarization continues to enhance the diversity and information richness, and strive to produce coherent and focused answers to users information need.