Chapter – 7 Conclusion and Future Work

The final and concluding part of the research is briefly summarized here. The work carried out in this research discusses two models for single document Text Summarization where the base line is Latent Semantic Analysis. The aim and objective being improvising the summary generated which is extractive. This better and precise summary has been evaluated using the Rouge Toolkit. The models used in this work have been developed using Python. In this chapter the first section discusses the conclusion of the two models and the second section deliberates on future work and enhancement that can be done on the work carried out so far.

7.1 Conclusion

The first model developed and discussed in Chapter 5 was a combination of LSA and the Naïve Bayes Classification. It essentially used LSA for choosing sentences with weights based on specific threshold given by the system. Further, using Naïve Bayes approach of Machine Learning, the model trains the classifier and predicts the summary that is built on the basis of calculation of Singular Value Decomposition (SVD). For the training of the model, two major steps involving the concepts of SVD - feature ranking and recursive feature elimination have been used. This model's performance was compared with that of Edmunson, LexRank, TextRank and SumBasic of the Rouge Toolkit. As has been clearly analyzed by applying the model on the standard DUC 2004 datasets, the proposed model generates a better summary and enhances values for the metrics of Accuracy, Precision and F-Score.

The second model which has been named as the Hybrid Model focuses on Machine Learning as well as Deep Learning techniques applied on the summary generated using LSA. The model uses - Self-Organizing Maps (SOM) is an unsupervised method and Artificial Neural Networks (ANN) is a supervised method. The work involves investigating the effect of adding mapped sentences from SOM visualization and re-training the inputs to ANN for ranking the

sentences. In each individual experiment of the hybrid model, a different mapping of SOM is added to the ANN as the input vector. The proposed Hybrid model uses Stochastic Gradient Descent which updates set of parameters in an iterative manner to minimize the cost function. In addition, using back-propagation, weight is being adjusted for the input vector. The empirical results show that the hybrid model using mapping clearly provides a comprehensive result and improves the F-score on ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4. This novel method has been implemented on different documents, which are publicly available like Opinosis, DUC 2004, DUC 2006, DUC 2007.

7.2 Comparison of both Models

Both the proposed models were compared using the same datasets and it was empirically found that the first Naïve Bayes based model works best when the document size is small i.e. approximately 100 words and we have documents which are classified into different pre-defined classes.

The second Hybrid model does not need pre-defined class labels for the documents and gives a good summary for large documents. Compared to the previous model this model does take more time to execute.

7.3 Future Enhancement

In this work the models are designed only for the English language. In addition to this they can be used only with well formatted documents with respect to the grammatical rules of the English language. The same can be carried out on other languages also. In extraction based summarization the important part of the process is the identification of important relevant sentences of text. There can be a number of different options which can possibly improve the summarization task. One option could be using the Boltzman machine and/or auto-encoder techniques of Deep Learning. Another option could be using fuzzy logic which is definitely a lucrative area of research.

It is very important that the work done be extended for multi document summarization with large datasets as well as domain specific datasets. Multi-document summarization has become an important task for assisting and interpreting text information in today's fast-growing information milieu. It is very difficult for human beings to manually summarize large documents of text.

Furthermore, there is a need to produce new, relevant and current information in the form of summary as new information and material keeps on getting added on the net daily. The summarizers which generate accurate summaries on-the-fly are going to be very popular soon.

For multi-document summarization systems, it is important to determine a coherent arrangement of the textual segments. A summary with improperly ordered sentences confuses the reader and degrades the quality or reliability of the summary itself. Further research may focus on this sentence ordering strategy.