IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 25-28
www.iosrjournals.org

Text Summarization for Telugu Document

Dr.M.Humera Khanam[1], S.Sravani[2]
[1]*Dept. of Computer Science and Engineering, SVU College of Engineering, Tirupati, India.*
[2]*Dept. of Computer Science and Engineering, SVU College of Engineering, Tirupati, India.*

Abstract: *Text Summarization is the process of reducing a text Document with a computer program in order to create a summary that retains the most important points of the original document. As the problem of information overload has grown, and as the quantity of data has increased, text summarization has became an important and challenging area in natural language processing. It is very difficult for human beings to manually summarize large documents of text. This paper presents a text summarization technique to summarize a text document by using Frequency based approach. This approach extracts the most important sentences of the text document which gives the meaningful information of the large document.*
Keywords: *Text Summarization, frequency based approach, Extraction*

I. Introduction

With the growing amount of data in the world, interest in the field of automatic summarization generation has been widely increasing. Text summarization involves reducing a text file into a passage or paragraph that conveys the main meaning of the text. The searching of important information from a large text file is very difficult job for the users thus to automatically extract the important information or summary of the text file[9].

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary[5]. The extractive summarization systems are typically based on techniques for sentence extraction and aim to cover the set of sentences that are most important for the overall understanding of a given document. Abstractive methods create an internal semantic representation to create a summary that is closer to what a human might generate. Such summary might contain words not explicitly present in original.

With the rapid growth of the World Wide Web (internet), information overload is becoming a problem for an increasing large number of people. Automatic summarization can be an indispensable solution to reduce the information overload problem on the web.

II. Approaches

Automatic summarization involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text.

a. Frequency based approach

Keyword Frequency:

The keywords are the top high frequency words in term sentence frequency. After cleaning the document calculate the frequency of each world. And which words have the highest frequency these words are called keywords[6]. The words score are chosen as keywords, based on this feature, any sentence in the document is scored by number of keywords it contains, where the sentence receives 0.1 score for each key word.

Stop Word Filtering:

In any document there will be many words that appear regularly but provide little or no extra meaning to the document. Words such as 'the', 'and', 'is' and 'on' are very frequent in the English language and most documents will contain many instances of them. These words are generally not very useful when searching; they are not normally what users are searching for when entering queries.

b. K-Means clustering

k-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

III. Proposed Work

In this project we summarized a large text document in to a passage which retains most important sentences of the document to give a main meaning of the text. To summarize the text we first tokenize the sentences. Now we clean the text document by removing full stops, stop words like conjunctions, adverbs etc.
Examples for stop words

1. మన(mana)
2. నేను(nEnu)
3. ఒక్క(okka)
4. కానీ(kaanee)
5. ఉన్న(unna)

After removing the stopwords from the text file count the frequency of each word in remaining text file and remove all the low frequency words from the text. Then select the keywords which have highest frequency. After that select the sentences which have keywords with highest frequency.

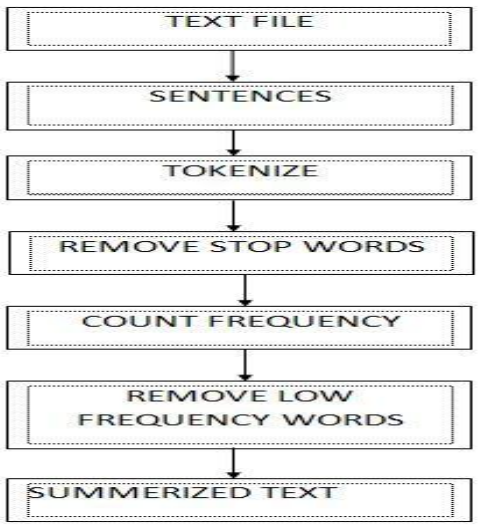
`In this technique, we first eliminate commonly occurring words and then find keywords according to the frequency of the occurrence of the word. This assumes that if a passage is given, more attention will be paid to the topic on which it is written, hence increasing the frequency of the occurrence of the word and words similar to it. Now we need to extract the lines in which extracted words occur since the other sentences wouldn't be as related to the topic as the ones containing the keywords would be. Thus, a summary is generated containing only useful sentences.

This technique retrieves important sentence emphasize on high information richness in the sentence as well as high Information retrieval. These related maximum sentence generated scores are clustered to generate the summary of the document. This takes into account facts such as the first few words of an article has more weights as compared to the rest. Secondly, it also takes into account the frequency of occurrence of keywords obtained in the this algorithm in a particular sentence. Higher the keyword count within a sentence, more is its relevance to the topic at hand.

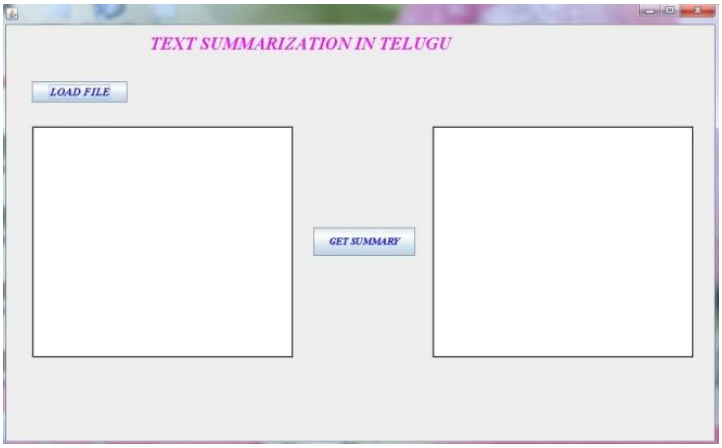
IV. Algoritham

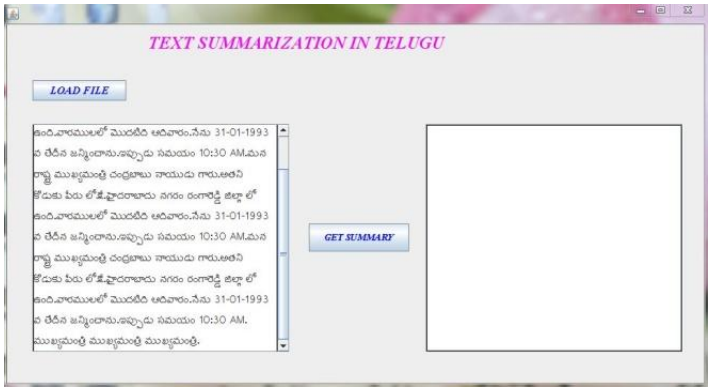
Step 1: read text file and perform tokenization using delimiters ". ,"
Step 2: create a list of stop words which
Step 3: Ignore stop words in text fie.
Step 4: Calculate frequency for remaining words in the text file by giving count.
Step 5: Consider words with maximum frequency count.
Step 6: Extract sentences from text file that contains words with maximum frequency words.
Step 7: output summary

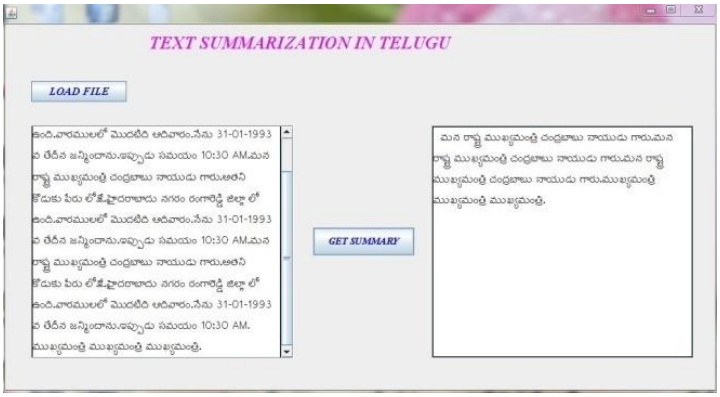
V. Block Diagram



VI. Results







VII. Conclusion

In frequency based technique obtained summary makes more meaning. But in k-means clustering due to out of order extraction, summary might not make sense.

The effective diversity based method combined with K-mean Clustering algorithm to generating summary of the document. The clustering algorithm is used as helping factor with the method for finding the most distinct ideas in the text. The results of the method supports that employing of multiple factors can help to find the diversity in the text because the isolation of all similar sentences in one group can solve a part of the redundancy problem among the document sentences and the other part of that problem is solved by the diversity based method.

In future work abstractive methods can be implemented. In abstractive method build an internal semantic representation and then use natural language generation techniques to create a summary.

References

[1] Inderjeet Mani, "Advances in Automatic TextSummarization", *MIT Press*, Cambridge, MA, USA, 1999.

[2] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summarizing textdocuments:sentence selection and evaluation metrics", *ACM SIGIR*, 1999, pp 121–128.

[3] E.H. Hovy and C.Y. Lin, "Automated Text Summarization in SUMMARIST", *Proceedings of the Workshop on Intelligent Text Summarization, ACL/EACL-97*. Madrid, Spain, 1997.

[4] J. Carbonell and J. Goldstein, "The use of MMR, diversitybased reranking for reordering documents and producing summaries," *ACM SIGIR*, 1998, pp. 335–336.

[5] Zha Hongyuan, "Generic Summarization and Key phrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering", *ACM*, 2002.

[6] John Conroy, Leary Dianne, "Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition", *ACM SGIR* ,2001.

[7] Daniel Marcu "From discourse structures to textsummaries" In *ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, 1997, pp 82–88.

[8] J. Pollock and A. Zamora "Automatic abstracting researchat chemical abstracts service", *JCICS*, 1975

[9] P. Kanerva, *Sparse distributed memory*, Cambridge, MA, USA: MIT Press, 1988.

[10] Y. Ko, et al., "Automatic text categorization using the importance of sentences," in Proceedings of the 19th International Conference on Computational Linguistics, Vol. 1, 2002, pp. 1-7.

[11] A. Kolcz, et al., "Summarization as feature selection for text categorization," in Proceedings of the 10th International Conference on Information and Knowledge Management, 2001, pp. 365-370.

[12] D. Shen, et al., "Web-page classification through summarization," in Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, 2004, pp. 242-249.

[13] A. McCallum and K. Nigam, "A comparison of event models for Naïve Bayes text classification," in Proceedings of AAAI Workshop on Learning for Text Categorization, 1998, pp. 41-48.

[14] R. R. Yager, "An extension of the naïve Bayesian classifier," Information Sciences, Vol. 176, 2006, pp. 577-588.

[15] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 143-151.

[16] I. Rahal and W. Perrizo, "An optimized approach for KNN text categorization using P-trees," in Proceedings of ACM Symposium on Applied Computing, 2004, pp. 613-617.

[17] E. Gabrilovich and S. Markovitch "Text categorization with many redundant features: using aggressive featureselection to make SVMs competitive with C4.5," in Proceedings of the 21st International Conference on MachineLearning, 2004, pp. 321-328.