GENERATING SUMMARY FOR A TELUGU TEXT DOCUMENT

Dr. M. Humera khanam[1]and S. Sravani[2]
[1]Department of Computer Science Engineering, SVU College of Engineering, Tirupai, India

[2]Department of Computer Science Engineering, SVU College of Engineering, Tirupai, 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. Text Summarization is a challenging problem these days. Due to the great amount of information we are provided with and development of Internet technologies, needs of producing summaries have become more important. Summarization is a very interesting and useful task that gives support to many other tasks as well as it takes advantage of the techniques developed for related Natural Language Processing tasks.
In this paper we propose a text summarization technique to summarize a Telugu document by using Frequency based approach and K-means clustering.

Keywords:Abstractive Summarization, Extractive Summarization, Frequency based approach, K-means clustering.

1. Introduction

Automatic summarization involves reduces 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 automatic extract the important information or summary of the text file.
This summary helps the users to reduce time instead of reading the whole text file and it provide quick Information from the large document. In today's world to Extract information from the World Wide Web is very easy. This extracted information is a huge text repository. Text Summarization methods are of two types 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.

1.**Proposed work:**

2.1 Frequency Based Approach:

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.The input to this method is a text document which contains telugu text. This text file is divided into sentences. Then, these sentences are tokenized in to words.

For example:

ఆంధ్రప్రదేశ్మరియుతెలంగాణరాష్ట్రాలఅధికారభాషతెలుగు .భారతదేశంలోతెలుగుమాతృభాషగామాట్లాడే 7.8 కోట్లజనాభాతోప్రాంతీయభాషలలోమొదటిస్థానంలోఉంది.

Sentences

- ఆంధ్రప్రదేశ్మరియుతెలంగాణరాష్ట్రాలఅధికారభాషతెలుగు.
- భారతదేశంలోతెలుగుమాతృభాషగామాట్లాడే 8.7 కోట్లజనాభాతోప్రాంతీయభాషలలోమొదటిస్థానంలోఉంది.

Tokenization

- ఆంధ్రప్రదేశ్|మరియు|తెలంగాణ|రాష్ట్రాల|అధికార|భాష|తెలుగు.
- భారత|దేశంలో|తెలుగు|మాతృభాషగా|మాట్లాడే|

  8.7|కోట్ల|జనాభాతో|ప్రాంతీయ|భాషలలో|మొదటి|స్థానంలో|ఉంది.

Now we clean the text document by removing stop words which are commonly occurring words and give less meaning in forming the sentence. To remove these stop words we have prepared a list of stop words in the Telugu language.

Examples for stop words

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

After removing the stop words from the text file count the frequency of each word in remaining text file. Then the words which have top high frequency are selected as keywords. Then summary is generated by extracting sentences which have these keywords. Thus the document is summarized in to a shorter form.

`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.

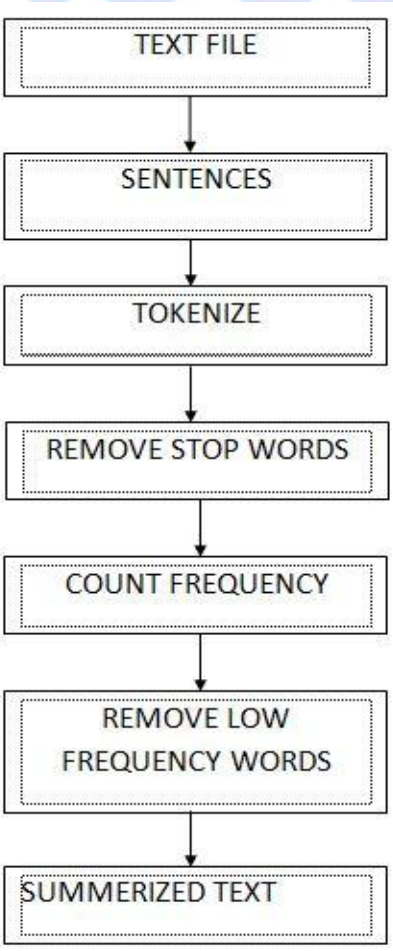
This technique retrieves important sentence emphasize on high information richness in the sentence as well as high Information retrieval. These related maximum sentence generated scores are clustered to generate the summary of the document. This takes into account facts such as the first few words of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

2.2 k-means Clustering

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided.

Here, we are using k means clustering algorithm for Text Summarization. It involves preprocessing, clustering and summary generation. The input to this method is a text document which contains Telugu text. In preprocessing step, text document is tokenized in to words. After tokenization, stop words which are commonly occurring words and which gives less meaning in forming a sentence are removed. To remove these stop words we have prepared a list of stop words in Telugu language. After removing stop words, frequency count is given to the remaining words in the document. Words with high frequency count are taken as keywords. Then clusters as formed with the sentences which contains these keywords taking keyword as a lable. Now we have to calculate sentence score. Keywords are given 0.1 as the feature value and all the remaining words are given 0. Sentence score is calculated by adding feature values of all words present in the sentence.Then we generate summary by selecting one sentence from each cluster based on sentence score.

1.Block diagram:

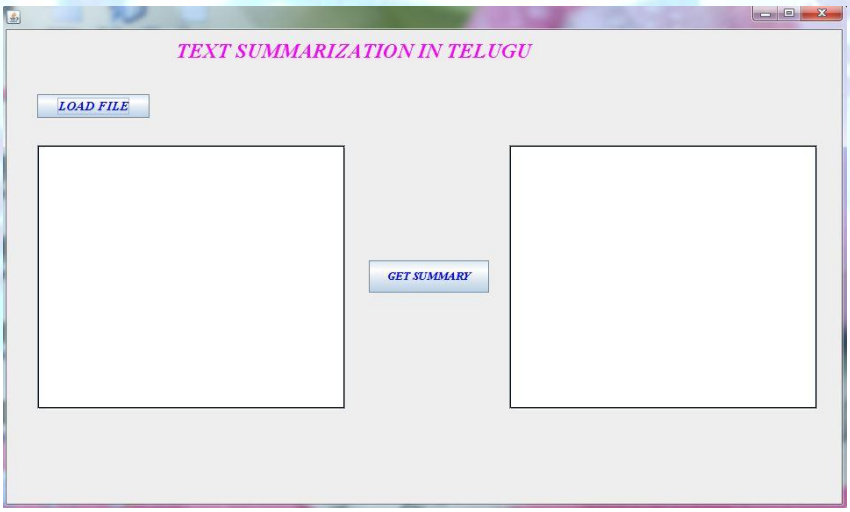


1.Algorithm:

Input: Telugu text document
Output: Summary for given document

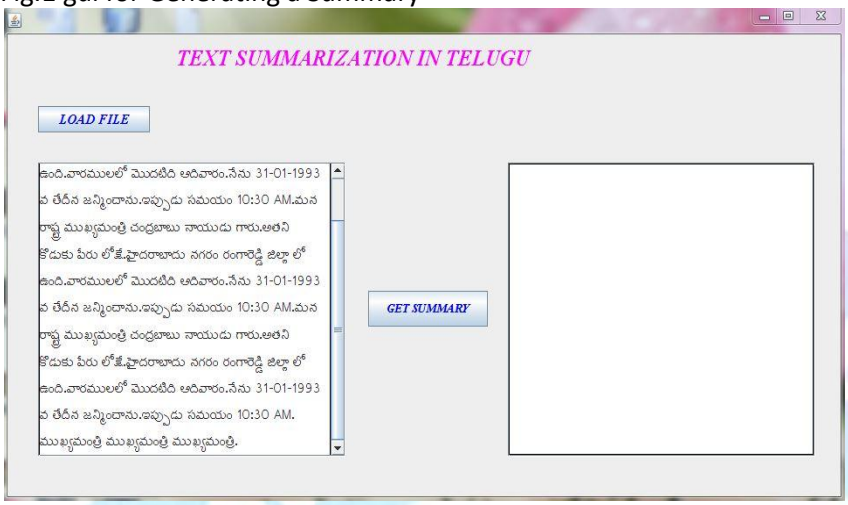
1. File(f)
2. Sentence segmentation(S).
3. Tokenize into words with delimiters " ,.!".
4. Prepare stop words list.
5. Remove stop words from given file.
6. Count frequency for remaining words in file.
   For i=1 to wordCount(f)
   currWord=word(i)
   Count(i)=0
   For j=1 to wordCount(f)
   If(currWord(i)==word(j)
   Count(i)=count(i)+1
   End
   End
7. Sort frequencies
8. Select high frequency words as keywords.
9. Extract sentences having keywords.
   For i=1 to S
   If(S_i contains any keyword)
   Extract S_i
   End
10. Generate summary.

1. Results



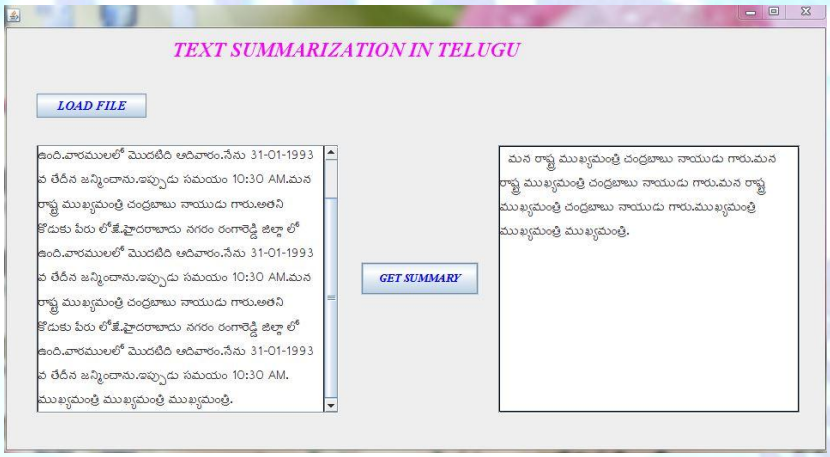
This figure shows the GUI interface for the text Summarization. Here load file allow user to load a Telugu text file.

Fig.1 gui for Generating a Summary



This figure shows a Telugu text file in the first column which has to be reduced in to a paragraph which gives the summary of the original document.

Fig.2 loading a Telugu text file



This figure shows the summary generated for the Telugu document. After loading the text file the summary of the document is generated by clicking Get Summary button.

Fig.3 generating summary

2. **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, diversitybasedreranking for reordering documents and producing summaries," *ACM SIGIR*, 1998, pp. 335–336.
5. ZhaHongyuan, "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.
18. Fang Chen, Kesong Han and Guilin Chen, "An Approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489-493, 2002.
19. C. Jaruskulchai and C. Kruengkrai, "Text Summarization Using Local and Global Properties", Proceedings of the IEEE/WIC International Conference on web Intelligence, 13-17 October. Halifax, Canada: IEEE Computer Society, 201-206, 2003.
20. M. Osborne. Using Maximum Entropy for Sentence Extraction. *In ACL Workshop on Text Summarization*, (2002).

21. Joel laroccaNeto, Alex A. Freitas and CelsoA.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach", Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215,2002.
22. NiladriChatterjee, Shiwali Mohan, "Extraction-Based Single-Document Summarization using Random Indexing" 19th IEEE International Conference on Tools with Artificial Intelligence.
23. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", in proceedings of the 10th European Conference on Machine Learning,1998,pp. 137-142.