International Journal of Research in Engineering, Science and Management
Volume-2, Issue-1, January-2019
www.ijresm.com | ISSN (Online): 2581-5792

Text Summarization Techniques Survey on Telugu and Foreign Languages

Sana Shashikanth[1], Sriram Sanghavi[2]
[1]*Assistant Professor, Department of Information Technology, JNTUH-CEJ, Karimnagar, India*
[2]*PG Student, Department of Computer Science and Engineering, JNTUH-CEJ, Karimnagar, India*

***Abstract*: Text summarization is the process of reducing a text document and creating a summary. Summaries are two types. Abstractive and Extractive summaries. An Extractive summary involve extracting relevant sentences from the source text in proper order. Abstractive summaries are in natural language regarding the source text that human might generate. Proposed work steps involves Pre-Processing sentences, Candidate sentence selection, Rank the sentences, removal of unnecessary sentences ,Extraction of sentences which are above threshold, thus summary is generated. Research lies in extraction of appropriate features and its application is to scoring of sentences. Words with high frequency are chosen to form a meaningful summarized text.**

***Keywords*: Extraction, Text Summarization, Frequency based approach, Summery generation.**

1. Introduction

Text Summarization is a technique in which large amount of text is compressed in such a way that actual meaning of data doesn't change. Text Summarization can be categorized into two types based on size of input, i.e., namely single document summarization, multiple document summarization. Single document summarizer means the summary is extracted from a single document, multiple document summarizer means the summary extracted from a multiple document. Generic document summarizer means generates summaries containing main topics of document, Query based document summarizer generate summaries containing searches that are related to given queries. Text Summarization methods are Abstractive and Extractive. 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 in the original text to form the summary. Abstractive methods create a summary that is closer to what a human might generate. With the rapid growth of the 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 Internet .This paper focuses on survey and performance analysis of automatic text summarizers for various foreign languages and Telugu language.

2. Text summarization techniques for foreign languages

Text summarization techniques for various foreign languages like

- Arabic
- Swedish
- English
- Turkish

A. Arabic

In 2006Mogda Fayek, Sobh Ibrahim and Darwish introduced a Bayesian approach for Arabic wrenching out text summarization. The main features used are positioning the sentences, length of sentences, length of paragraphs. The system performance can be increased by combining weight of sentence, length of sentence and position of sentence. The system performance slightly changed on using the position of the paragraph and sentence paragraph, the average result is nearly 68%. In 2012, Tarek and Fatma El-Ghannam presented a best summarization algorithm for Arabic based on wrenching out summarization approach. In a document, important keys are recognized by using the various combinations of linguistic and statistical features. To rand the sentences extraction algorithm uses key phrases. The average result is nearly 66%.

B. Swedish

Gustavsson and Arne Jhonsson both presented results from evaluations of an automatic text summarization technique. This experiment makes use of both Random Indexing and Page Rank. Types of text used in this area of two types: newspaper text and government text. The experiment analysis show that types of texts affects the performance. Combining Random Indexing and Page Rank on government texts gives best results.

C. English

H.p.Luhn is called as father of automatic text summarization. At first he created an abstract of English technical literature to get fast and accurate information of technical papers. The main goal of Luhn was to save readers time in getting information from articles and reports. The document was converted into machine readable format and it was scanned by IBM 704 machine. The statistical information is taken from the processed text. Word frequency and distribution was used for computing

relative measures. Frequency of each word is calculated at first and later for sentences.

D. Turkish

In 2010 Mucahid, Ilyas Ciecekli and Celal Cigir presented the generic text summarization technique for Turkish language. In this, to extract sentences text content is cover with less redundancy and many more features are used such as centrality, frequency of terms label similarity and position of sentences. The rank of the each sentence is counted using a score function which defines its featured values and the feature weights. Using machine learning, feature weights are learned with the co-operation of human generated summaries. By comparing manual summary with output summary performance is evaluated from Turkish text documents.

3. Text summarization techniques for Telugu languages

Summarization for Telugu documents involves various techniques. These techniques helps in reducing the size of the document that results a short set of words that gives main meaning of the text. They are as follows-

- K-Means Clustering
- Frequency based approach

A. K-Means clustering

K-Means algorithm is for finding similar group of data by forming various clusters. It is an Iterative algorithm which follows two steps.

- Cluster Assignment.
- Move Centroid.

This algorithm goal is to divide M observations into K clusters in which each observation is a part of a cluster with nearest mean.

B. Frequency based approach

1) Frequency of keywords

In this we are going to calculate the frequency of each word. The words with maximum frequency are called Keywords. Based on keywords, Score is awarded to each sentence. For each keyword the sentence gets 0.1 score.

2) Filtering of stop words

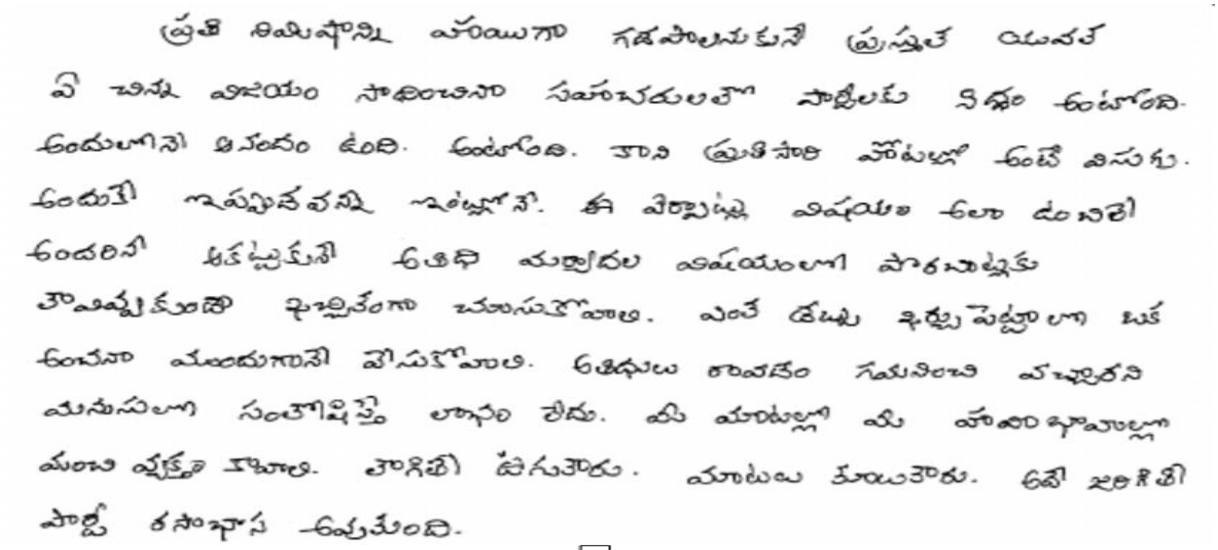
In any document there are words without having any meaning used as Prepositions such as ON, AND, WITH, AS Etc., used very frequently in English language. Those words are not useful for what users are searching for, while executing queries.

4. Research work

In this the text document is taken. The sentences from the text are read and are tokenized. The document is cleaned by removing the stop words like full stops, commas etc. Stop words examples are

- నేను(nenu)
- ఒక్క(okka)
- ఉనన(unna)
- కానీ(kaanee)

After removing stop words, frequency of each word is calculated. Words with low frequency are exempted from the



యువత ఏ చిన్న విజయం సదించిన పర్త్యే ల కు సిద్ధం అంతుంది.ప్రతిసారి హొతెల్ లొ విసుగు చెంది ఇంత్లొనె జరుపుథుంది.అథిది మర్యదలథొ థక్కువ దబ్బులు కర్చు పెత్తి మంచి హవబవం లు వ్యక్థం చెయలి.పర్త్యే రసబస కకుంద చుదాలి.

Fig. 1. Result

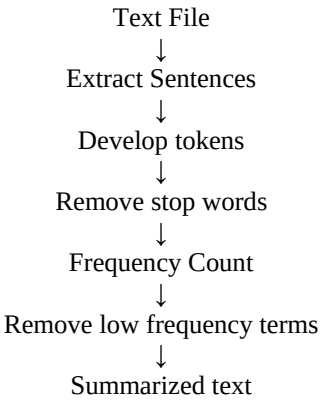
summary. Words with high frequency are chosen to form a meaningful summarized document. Summarized document gives the main meaning of the text. So reduces the reader's time and helps to understand the content of the document easily.

A. Algorithm

Steps

- Read the text file from the user and extract sentences.
- Develop the tokens from the sentences.
- After tokenization remove the stop words.
- Count the frequency of each word in the document.
- Remove the low frequency words from the document and consider the words with maximum frequency count.
- Summarized document.

B. Flowchart



5. Results

The result is shown in Fig. 1.

6. Conclusion

In this paper, a brief summary of automatic text summarization techniques for various foreign and Telugu languages has been described. We noticed that good work has been done for various foreign languages but summarization for Telugu language is still lacking. In K-Means clustering because of order extraction summary might not be meaningful where as in frequency based technique resulted summary makes more meaning .We can conclude that various combinations work differently for various types of content. In future we are aiming at to use more features for extracting Telugu sentences. Hence, it is challenging to create a single summarizer for various types of Telugu content.

References

[1] El-Shishtawy, Tarek, and Fatma El-Ghannam, "Keyphrase based Arabic summarizer (KPAS)." In 8th International Conference on Informatics and Systems (INFOS), pp. NLP-7, IEEE, 2012.
[2] M. Kutlu, C. Cigir, and I. Cicekli, "Generic text summarization for Turkish." The Computer Journal, vol. 53, no. 8, pp.1315-1323, 2010.
[3] Gustavsson, Pär, and Arne Jönsson. "Text summarization using random indexing and pagerank." Proceedings of the third Swedish Language Technology Conference (SLTC-2010), Linköping, Sweden. 2010.
[4] Hans Peter Luhn. "The automatic creation of literature abstracts," IBM Journal of research and development, 2(2):159–165, 1958.
[5] Prachi Shah, Nikita P. Desai. "A survey of automatic text summarization techniques for Indian and foreign languages". International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)-2016.
[6] M. Humera Khanam, and S. Sravani. "Text Summarization for Telugu Document". IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), pp. 25-28.