TEXT SUMMARIZER

MERIN ANNA MATHEW, NOBY BENNY, NKUMUM DEMAS MERIKI,
RIYA ELIZABATH VINOTH, Dr.ARUN KS*
*Department of CSE, Amal Jyothi College of Engineering,
Kanjirapally, Kerala, India,

Abstract

In this present era, where huge quantity of information is generating on the internet day by day. So it is necessary to provide the better mechanism to extract the useful information fast and most effectively. Text summarization is one of the methods of identifying the important meaningful information in a document or set related document and compressing them into a shorter version preserving its overall meanings. It reduces the time required for reading whole document and also it space problem that is needed for storing large amount of data.Automatic text summarization problem has two sub-problems that is single document and multiple documents. In single document the single document is taken as the input and summarized information is extracted from that particular single document. In Multiple document the multiple documents of single topic is taken as an input and the output which is generated should be related to that topic.

I. Introduction

As the amount of information on the web is increasing rapidly day by day in different formats such as text, video, images. It has become difficult for individual to find relevant information of his interest. Suppose user queries for information on the internet he may get thousands of result documents which may not necessarily relevant to his concern. To find appropriate information, a user needs to search through the entire documents this causes information overload problem which leads to wastage of time and efforts. To deal with this dilemma, automatic text summarization plays a vital role. Text summarization now,has emerged as untimely tool for our daily day purposes. Manual text summarization is a very big deal.So, this gives a better mechanism to get the useful information accurately and effectively. Text summarization is one of the methods of identifying the important meaningful information in a document or set related document and compressing them into a shorter version preserving its overall meanings. It is currently receiving a great deal of attention and now indicates that both human effort and bias may be eliminated by this process. Its always a humanly thing to accomplish things ,with reduced time.This also had effect,on the manner of reading large documents.The main characteristics that we all need while reading summaries,are :If the correct meaning is emphasized when,converting large document into a summary. And,if it has some effective time reducement while reading so.Thus,effective summary making

has become a need of time. Collecting the diverse topics,keeping up the redundancy,time management are the factors that might be kept in mind in putting together a text summarizer. In Automatic Text summarization has two modes namely :Abstractive and Extractive text summarization. An extractive text summarization ideology is extracting the important text from the input document. An extractive text summarization approach uses semantic or data features for selecting useful informative sentence. An Abstractive text summarization will try to understand the input file or original file and re-generate the output in few words by identifying the main concept of the input file. In many research papers they have mentioned that extractive text summarization is sentence ranking. Extractive text summarization is divided in two phases: 1) Pre-processing 2) Processing. In this paper we are explaining extractive text summarization on single document 1)Pre-Processing refers to the structured representation of the original document. It includes: a) Identification of the Sentence boundary. In English, which is identified with a dot at the end. b) Eliminating stop-words —Common words with no semantics and which do not add up relevant information to the task are eliminated. c) Stemming— The purpose is to obtain the radix of each word, emphasizing its semantics. 2)Processing step features, the factors that influence the relevance of sentences and calculated and then weights being assigned to those using weight learning method. Final score of each sentence is determined then by, Featureweight equation. Top ranked sentences are eventually selected for final summary. This paper describes some exploratory research on automatic methods of obtaining abstracts. The system outlined here begins with the document in machine-readable form and proceeds by means of a programmed sampling process comparable to the scanning a human reader would do. However, instead of sampling at random, as a reader normally does when scanning, our method selects those among all the sentences of an article that are the most equivalent representation of relevant information. These key sentences are then enumerated to serve as clues for judging the character of the article.

II. RELATED WORK

Interest in automatic text summarization, arose during the early fifties. An important paper of those days is the one published in1958,which suggested to weigh the sentences as a function of high frequency words, disregarding the other set of high frequency common words. Automatic text summarization model back in sixties, in addition to the standard keyword method of frequency depending weights, also used the following three methods for determining the sentence weights:

Cue Method: This is based on the relevance of a sentence by particularly computing the existence or truancy of certain cue words in the cue dictionary.

Title Method: Sentence weight computed as a function that sums all the content words present in the title and sub headings of a text.

Location Method: Based on the assumption that sentences that hold the initial position of both text and individual paragraphs have a higher degree of relevance. The results also showed, that the best way to match-up both the automatic and human-made extracts can be achieved using a

combination of these three latter methods.

The Trainable Document Summarizer in
1995, performs sentence extracting task, based on a number of weighting heuristics. Following features were used and evaluated: 1. Sentence Length Cut-Off Method: Sentences with less than a pre-specified number of words are not included in the extract. 2. Fixed-Phrase Method: Sentences having certain cue words and phrases cannot be eliminated.3. Paragraph Feature: Basically equivalent to Location Method features. 4. Thematic Word Feature: The most frequent words are itself mentioned as 'the thematic words'.Sentence scores are used to count these frequencies 5. Uppercase Word Feature: Uppercase words with certain obvious exceptions are also treated as thematic words,as well.

Natural Language Processing arose as the centuary's limelight, since the time human languages are interpreted at a large scale from each other. Summarization is feather in the cap of NLP reasearch works which concentrates on providing meaningful summary from text documents employing various NLP tools and techniques. Since amount of data information used across the digital world rises day by day,it is highly essential to have available automatic summarization techniques. Research works are being carried out in the area of Extractive summarization. Even though more works come under Extractive method which is quite popular , meaningful summary attained using the Abstractive summary techniques are also welcomed but making the process more complex.

III. PROPOSED SYSTEM

Automatic text summarization is a complex task which contains many sub-tasks in it. Every subtask has an ability to get good quality summaries. The important part in extractive text summarization is identifying necessary paragraphs from the given document. In this work we proposed extractive based text summarization by using statistical novel approach based on the sentences ranking the sentences are selected by the summarizer. The proposed model improves the accuracy when compared traditional approach. We are taking input as text file .txt. Extractive summarizers aim at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary.
Machine Learning approach From the set of training documents and their extractive summaries, the summarization process can be modeled as a classification problem.In which, the sentences are classified as both suitable and non-suitable sentences for the summary,based on the features that they show. This classification probabilities are learnt significantly from the training input, using the Bayes' rule:

$$P(s \subset S|F1, F2...FN) =$$

$$P(F1, F2...FN|s \subset S) * P(s \subset S)/P(F1, F2...FN)$$

where s is a sentence from the document collection, F1, F2. . . FN are features used in classification. S is the summary to be generated, and P (s< S | F1, F2, ..., FN) is the probability

that sentence s will be chosen to form the summary given that it possesses features F1,F2. . . FN. The proposed model can be put into diagram as follows,

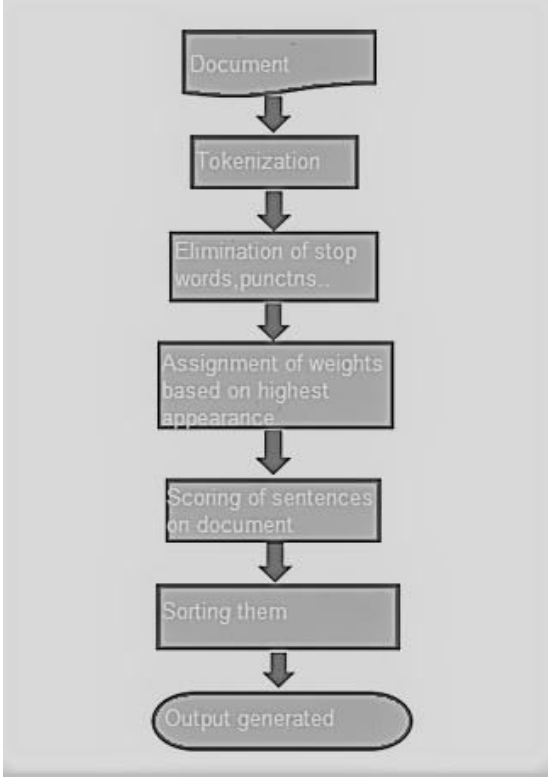

Figure 1: *Flowchart.*

As soon as the search image,the search folder and the result folder are submitted by the user the analysis process is formed.The search image undergoes detection phase and encoding of the face thus detected is generated.Meanwhile the application traverses through all the subdirectories of the search folder selected by the user,generating encodings of each of the image file.These encodings are compared with that of the search image using the euclidean distance.A suitable tolerance parameter is used to improve the accuracy.

IV. RESULTS

In the extractive summarization, the summarizer takes input as text file and tokenization of an input text is done in-order to remove find the terms of the text. Then stop words are removed in order to filter the text. And finally, part-of- speech tag is added to each token. Step 1: After adding the parts-of-speech tag to tokens or terms each individual weight are assigned to the tokens. The term weight is calculated as follows:

$$W_t = freqofterm/totalnoofterms$$

Step 2: Now maximum weight of the token is considered after finding maximum weight.The weighted frequency of the document is calculated as follows:

$$W_{tf} = freqofaterm/maxfreqoftheterm$$

Step 3: In this step, the frequencies are connecting in place of corresponding words in sentence and

sum of it is found. The ranks are found based on the weighted frequency. The sentences are sorted based on their Weighted frequency ranks like highest rank to lowest.The sentences are arranged in descending order.

Finally, summarizer will extract sentences which rank is highest form the document and then sentences which are extracted.

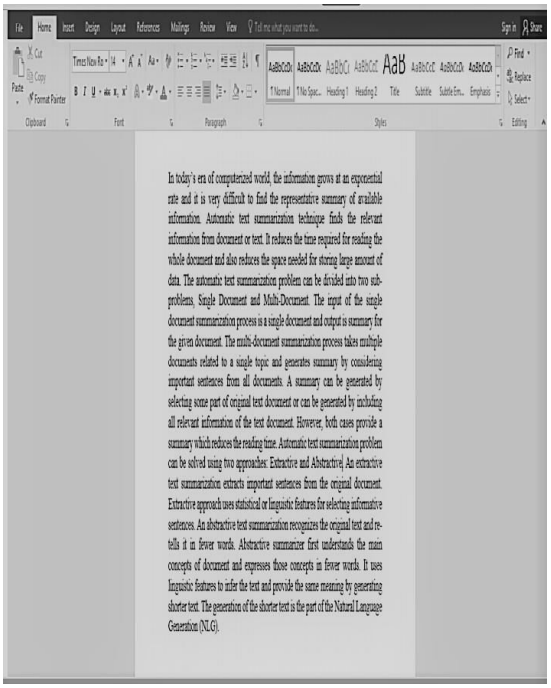

Figure 2: *Input.*

Figure 3: *Tokenization.*

Figure 4: *Largest Score.*

Figure 5: *Frequency Table.*

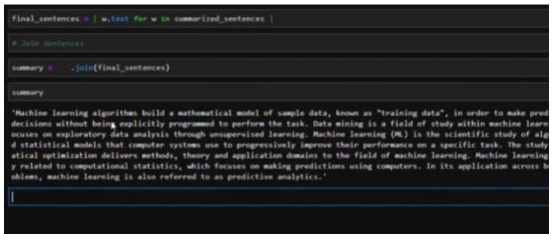

Figure 6: *Join Sentance.*

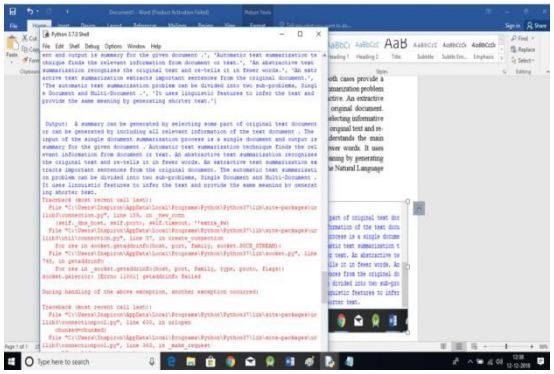

Figure 7: *Output generated.*

V. CONCLUSION

In this work we proposed extractive based text summarization by using statistical novel approach based on the sentences ranking the sentences are selected by the summarizer.The sentences which are extracted are produced as a summarized text. The proposed model improves the accuracy when compared traditional approach. The biggest challenge for text summarization is that their input content come from a number of textual and semi-structured sources, including databases and web pages which is uneven.

The deal of an effective text summarization software is the production of effective summary in less time and with least redundancy. Summaries can be evaluated using both intrinsic or extrinsic means. While intrinsic methods are an attempt to measure summary quality with the help of human evaluation and extrinsic methods measures the same through a task based performance measure.Retrieval oriented is one such.For texts with multiple area of significance, the generated summary can be or can be not balanced. Deciding proper weights of individual features is another important quality of final summary is. We should invest more time in deciding feature weights. There is, of course, the chance that an author's style of writing deviates from the average so that, to an extent this might cause the method to select sentences of inferior significance also.

VI. REFERENCES

[1] Nenkova, A.(2011).“Automatic summarization, Foundations and Trends in Information Retrieval”,5(2),103-233

[2] Gupta,V and Lehal,G.s (2010). “A survey of text summarization extractive techniques.” Journal of Emerging Technologies in Web Intelligence,2(3),258-268

[3] Goldstein.J,carbonell.J,Kantrowitzt. M(1998).“Multiple document summarization by sentence Extraction”40-48

[4] Weigo Fan, Linda Wallace, Stephanie Rich and Zhongju Zang,: “Tapping the power of text mining”, Journal of ACM,Blacksburg 2005.

[5] Baxendale,P.(1958). “Machine-made index for technical literature” –an experiment. IBM Journal of Research developement354-361

[6] Vishal Gupta,G.s. Lehal. “A survey of text mining techniques and applications”, Jour-nal of Emerging Technologies in Web intelligence,VOL 1,NO 1,6076,August 2009