IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

Extractive Text Summarization

Vivek S. Bhore[1], Pratik Bondare[2], Rutik D. Gawande[3], Vrushabh V. Guntiwar[4], Priti V. Kale[5]
Project Group Leader, Department of Information Technology [1]
Project Group Member, Department of Information Technology[2,3,4]
Project Guide, Department of Information Technology [5]
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India

Abstract: *In this fast paced technological era, where huge quantity of information is generating on the internet day by day. Since the dotcom bubble burst back in 2000, technology has radically transformed our societies. So, it is necessary to provide the better mechanism to extract the useful information fast and most effectively. Automatic 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 reduces space that is needed for storing large amount of data. Automatic Text summarization has two approaches 1) Abstractive text summarization and 2) Extractive text summarization. In extractive text summarization only important information or sentence are extracted from the given text file or original document. Here we will discuss on extractive text summarization using sentence scoring and sentence ranking method.*

Keywords: Dotcom Bubble; Automatic Text Summarization; Abstractive Text Summarization; Extractive Text Summarization; Sentence Scoring; Sentence Ranking

I. INTRODUCTION

Data mining, also known as knowledge discovery in the data, is a process of uncovering patterns and other valuable information from the large data sets. Automatic text summarization is one of the Data mining techniques. There are two main types of automatic text summarization, as shown in the below diagram.

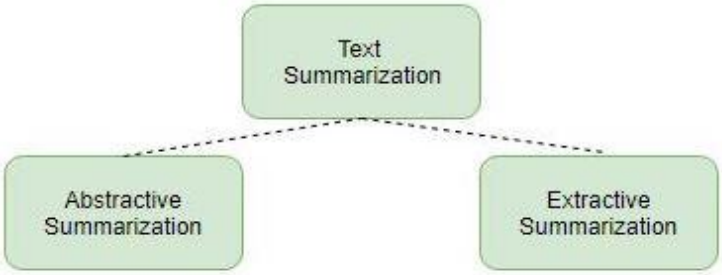

Figure 1: Types of Text Summarization.

In this article we will focus on extractive text summarization using sentence scoring method and sentence ranking method. An extractive text summarization approach uses linguistic or statistical features for selecting useful informative sentence. . 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. In this paper, we will see single document summarization using extractive method. Extractive text summarization using sentence ranking involves two phases: Pre-processing and Processing whereas the text summarization using sentence scoring involves four phases: - Pre-Processing, Sentence Scoring, Sentence Ranking, Summary Extraction.

To maintain proper flow of the content this paper is organized in various chapters; Second chapter is related works which includes several past documents, manuals, analysis papers which are already published and are associated with extractive text summarization. Third chapter will focus on the proposed approach for summarization, Fourth chapter will focus on the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

method which is involved during implementation of this techniques, and the last chapter will focus on the conclusion that can be drawn out after this analysis.

II. Related Work

This section describes already existing studies from several documents, manuals, analysis papers that have been conducted on Text summarization. In earlier researches, Summarization was done on scientific documents based on the proposed features like phrase frequency, word (Luhn, 1958) [1], key phrases (Edmundson, 1969) [5] and position in the text (Baxendale, 1958) [3].

In 1958 **Luhn's** has researched on text summarization he extracted important sentence by using the position of text. His finding showed that frequency of words in sentences has more significance in the final outcome. The methods proposed by Luhn are still effective even though they are over 50 years old. He also proposed removal of stop words, stemming i.e. converting the words to their root form. The words are given a hierarchy and each word's significance is described by its index. This will then calculate the number of time that particular word occurs in the sentence and then it is ranked according to that [1].

In 1969, **Edmundson** has done research on extracted summarization in this he extracted important sentence by using two features position and word frequency importance were taken from the previous works. He used the word frequency and word position feature. He also gave us two new features, cue words and skeleton. The sentences were scored basing upon these features which were then extracted for summarization [5].

In 1958, **Baxendale** has done his research at IBM on Extractive summarization. He extracted important sentence by using the position of text. In his study on over 200 paragraphs found that, in over 85% of those paragraphs the topic of the para would appear in the first sentences itself. And in 7% of the paragraphs the topic would appear in the last sentence. By this he came to a conclusion, that most of the times the topic appears in either the first or last sentence of the paragraph [3].

Fang Chen et al in their work observed 3 features. The Sentence location feature meant that most of the times the beginning and the end of the sentences would contain the useful matter. The second one is the paragraph location feature which is same as the sentence location feature. The third feature is the sentence length feature where the sentences that are too long or too short are not featured in the summary. The threshold for the number of words can be preset [4].

Hongyan Jing observed from his work that removal of irrelevant phrases like prepositional phrases, clauses, to infinitives, gerunds from sentences was of prime importance as they don't have any significance in the summarization process [2]

III. Approach

In this proposed approach, we are using extractive method to get summary of given input. We are taking input as text file .txt.

1. Firstly, the file which is given as input is tokenized in order to get tokens of the terms.
2. The stop words are removed from the text after tokenization. The words which are remained are considered as a key word
3. The key words are taken as an input for that we are attaching a part of tag to each key word.
4. After completing this pre-processing step we are calculating frequency of each keyword like how frequently that key word has occurred from this maximum frequency of the keyword is taken.
5. Now weighted frequency of the word is calculated by dividing frequency of the keywords by maximum frequency of the key words.
6. In this step we are calculating the sum of weighted frequencies.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

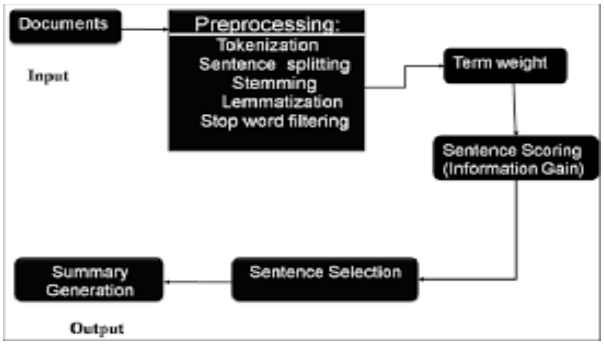

Figure 2: Model for text summarization using Information Gain Method.

Finally, summarizer will extract the high weighted frequency sentences and the extracted sentences are converted into summary format.

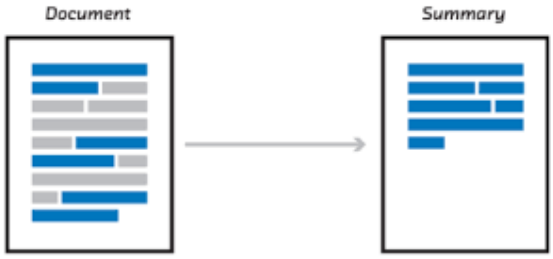

Figure 3: Summarizer.

IV. METHODOLOGY

Extractive text summarization is selecting the most relevant sentences of the text. This method consists of four phases, they are:

1. Pre-processing
2. Sentence scoring
3. Sentence ranking
4. Summary Extraction.

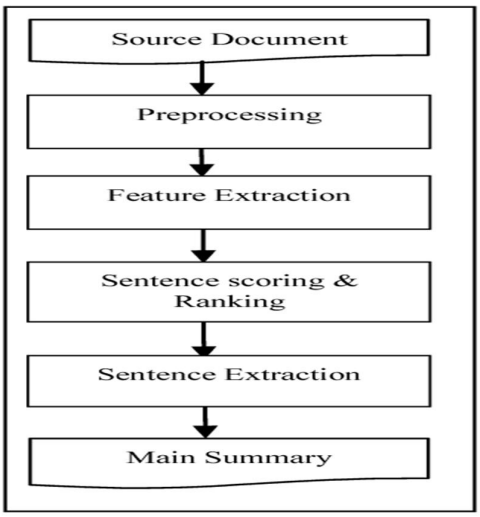

Figure 4: Phases involved in Extractive Summarization.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

4.1 Phase I: Pre-processing of input document

The phase of pre-processing involves chopping the paragraph into words. This phase involves four stages.

1. Sentence segmentation
2. Tokenization
3. Stop word Removal
4. Stemming.

In each stage the document undergoes different changes. The changes are explained below

4.1.1 Sentence Segmentation of paragraph in the document:

Sentence Segmentation is the process of breaking down/segmentation the given text document into sentences al. In this system sentence is segmented by identifying the boundary of sentence which ends with period symbol (.), question mark (?), exclamatory mark (!) and the total number of sentences present in the document are also identified.

4.1.2 Tokenization of segmented sentences :

Tokenization is the process of breaking down the sentences into words. Tokenization is done by identifying the spaces and special symbols between the words. In this process frequency of each word is calculated and stored for further processing.

4.1.3 Stop Word Removal from the list of words:

Stop words are the words that do carry as important meaning as by keywords. These words are identified by supplying a list of words with less importance to the system. The system compares these stop words with the tokenized words obtained from previous phase. These stop words are then disposed as they can interfere and influence the summary that will be generated at the end.

4.1.4 Stemming A word can be found in different forms in the same document:

These words have to be converted to their root form for simplicity. This process is known as Stemming. An algorithm is used to transform words to their root forms. In this system, Porter's stemmer method is used to turn a word into its root form using a predefined suffix list. Finally, frequency of each is word is calculated a retained for next phase.

4.2 Phase II: Sentence scoring After phase:

The input document is segmented into collection of words in which each word has its individual frequency. In phase 2 the sentences are ranked based on seven important features:

1. Frequency
2. Sentence Position
3. Cue words
4. Similarity with the Title.
5. Sentence length.
6. Proper noun.
7. Sentence reduction.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

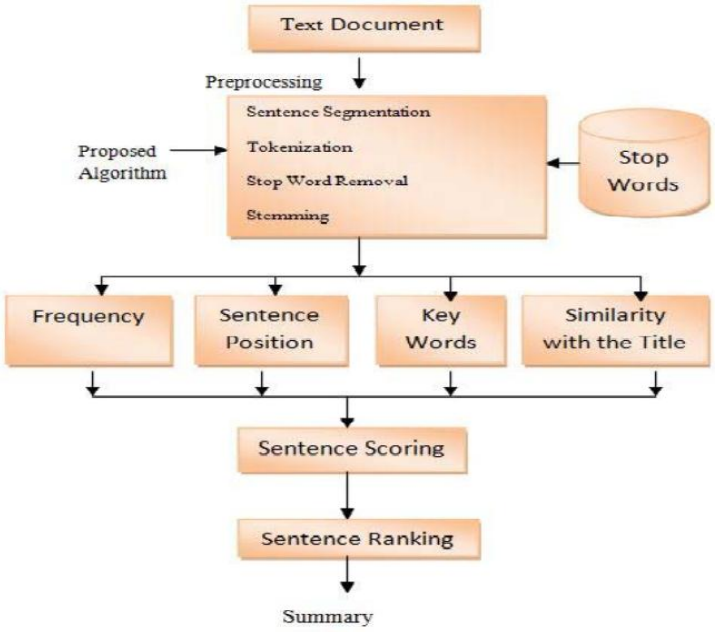

Figure 5: Text Summarization architecture.

4.2.1 Frequency

Frequency is the number of times a word occurs in a document. If a word's frequency in a document is high, then it can be said that this word has a significant effect on the content of the document. Salient sentences/words are those sentences/words that occur repeatedly. The frequently occurring word increases the score of sentences they are in. The most common measure widely used to calculate the word frequency is TF (Term frequency) IDF (Inverse document frequency). The total frequency value of a sentence is calculated by summing up the frequency of every word in the document.

4.2.2 Sentence Position Value

It depends on our requirement whether important sentences are located at certain position in text or in paragraph. Sentences in the beginning define the theme of the document whereas sentences in the end conclude or summarize the document. The positional value of a sentence is calculated by assigning the highest score value to the first sentence and the last sentence of the document. Second highest score value is assigned to the second sentence from starting and second last sentence of the document. Remaining sentences are assigned a score value of zero.

4.2.3 Cue Words

Cue words are the important words in a document. These Cue words are given as input from the user. If a sentence contains these Cue words then score value one is assigned to the sentence, otherwise the score value of the sentence will be zero.

4.2.4 Similarity with the Title

The words in the title and heading of a document that reappear in sentences are directly related to summarization. These words are considered for summarization as the have some extra weight in them. If a sentence contains words in title and header then score value one is assigned to that sentence, otherwise score value is zero for the sentence.

4.2.5 Sentence Length

The length of the sentence resembles the importance of sentence in summarization. Generally, sentences that are very long and very short are not suitable for summary. Sentences that are very long will have unnecessary information which is

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

not useful for summarization of document. Whereas, sentences that are too short do not give much of information about the document.

V. CONCLUSION

This paper discusses the simple and easy extractive technique of text summarization. The important part in extractive text summarization is identifying necessary paragraphs from the given document. This paper proposes extractive based text summarization by calculating scores of word and sentence features and then using this scores for ranking the sentences by using statistical novel approach based on the sentences ranking. The sentences which are extracted are produced as a summarized text by the summarizer.

The increasing progression of the Internet has made a huge amount of information available. It is very difficult for humans to summarize huge amounts of text. So, there is an immense need for automatic summarization systems in this age of information excess. Due to the rapid growth of knowledge and use of Internet, there is information overload. This problem can be solved, if there are robust text summarizers which produces a summary of document to help user. Hence, there is a necessity to develop system where a user can efficiently retrieve and get a summarized document. One potential solution is to summarize a document using either extractive or abstractive methods. The text summarization by extractive is easier to build.

REFERENCES

[1]. Luhn, H (1958). "The automatic creation of literature abstracts". IBM Journal of Research Development, number 2, pages 159-165, 1958.
[2]. Hongyan Jing, "Sentence Reduction for Automatic Text Summarization", pages 310-315, 2000.
[3]. P.B. Baxendale. "Man-made index for technical literature - An experiment". pages 354-361, 1958.
[4]. Fang Chen, Kesong Han and Guilin Chen, "An Approach to Sentence Selection Based Text Summarization", Volume: l, pages 489- 493, 2002
[5]. H. P. Edmundson. "New Methods in Automatic Extracting. Journal of. ACM", 16(2):264-285, 1969
[6]. T. Sri Rama Raju and Bhargav Allarpu, "Text Summarization using Sentence Scoring Method", Volume: 04, pages 1777-1779, Dept. Of CSE Engineering, GITAM University, Andhra Pradesh, India, April 2017
[7]. J.N.Madhuri and Ganesh Kumar.R, "Extractive Text Summarization Using Sentence Ranking", Dept. of Computer science and Engineering CHRIST, Banglore, India, IEEE 2019.
[8]. A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
[9]. J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
[10]. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.