International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

Text Summarization using Sentence Scoring Method

T. Sri Rama Raju, Bhargav Allarpu

Student, Dept. Of CSE Engineering, GITAM University, Andhra Pradesh, India
Student, Dept. Of CSE Engineering, GITAM University, Andhra Pradesh, India

---

Abstract - *In this project an automated text summarization tool has been developed using Sentence Scoring Method which involves finding the frequent terms, sentence ranking etc. Summary is extracted from the list of top ranked sentences. The size of summary can be specified by the user.*

***Key Words*: Text summarization; sentence scoring; sentence ranking.**

1. INTRODUCTION

Text summarization is a method used to generate summaries of text documents by extracting important information from the summarized document. It is used extensively in summarizing the search engine results, providing brief version of large documents in which abstracts are not present.

Text summarization involves two types, abstractive and extractive. In ABSTRACTIVE it considers human knowledge and understands the text in order to generate the summary whereas in EXTRACTIVE the important sentence is picked from the text as a summary. In this paper, we will see single document summarization using EXTRACTIVE method. However multiple documents can also be summarized by choosing the proper integration algorithms.

The text summarization using sentence scoring involves four phases: - Pre-Processing, Sentence Scoring, Sentence Ranking, Summary Extraction. The text document is selected and data in it is segmented into sentences and tokens during the pre-processing stage. Each sentence is evaluated in sentence scoring by considering the linear combination of multiple parameters like frequency, sentence position, cue words, similarity with title, sentence length and proper noun. The sentences are ranked with respect to the scores. The summary is generated by selecting the required number of sentences with highest rank and it is made sure that no two consecutive sentences are selected.

2. Related Work

In this section, we will discuss certain other research studies that have been conducted on Text summarization

LUHN's work on text summarization 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].

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].

BAXENDALE 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].

EDMUNDSON typical structure that produces extract. 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].

3.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

3.1 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

3.1.1 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.

3.1.2 Tokenization of segmented sentences

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

3.1.3 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.

3.1.4 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.

3.2 Phase II: Sentence scoring

After phase 1 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.

3.2.1 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.

3.2.2 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.

3.2.3 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.

3.2.4 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.

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

3.2.6 Proper Noun

Proper nouns play an important role in summarization. It gives information regarding, to whom or to what the author is referring. Roles played by individuals or locations description will be different more number of times in a document.

3.2.7 Sentence reduction

Sentence reduction is the method of removing irrelevant phrases like prepositional phrases, clauses, to infinitives, or gerunds from sentences. The goal is to identify less important phrases in a sentence using reduction decisions. The reduction decisions are based on syntactic knowledge, context, and probabilities computed from corpus analysis.

The final score is a Linear Combination of frequency, Sentence positional value, Cue Words, Similarity with the title of the document, Sentence length and Proper noun.

3.3 Phase III: Sentence Ranking

After each sentence is scored they are arranged in descending order of their score value i.e. the sentence whose score value is highest is in top position and the sentence whose score value is lowest is in bottom position

3.4 Phase IV: Summary Extraction

After ranking the sentences based on their total score the summary is produced selecting certain number of top ranked sentences where the number of sentences required is provided by the user. For the reader's convenience, the selected sentences in the summary are reordered according to their original positions in the document.

3. CONCLUSIONS

This paper discusses the simple and easy extractive technique of text summarization. This project is done mostly with java and any number of extensions for scoring techniques can easily be added.

REFERENCES

[1] H.P. Luhn, "The automatic creation of literature abstracts", in IBM Journal of Research Development, volume 2, 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,pp.489- 493,2002.

[5] H. P. Edmundson. "New Methods in Automatic Extracting. Journal of. ACM", 16(2):264-285, 1969.

[6] Khan Atif, Salim Naomie, "A review on abstractive summarization Methods", Journal of Theoretical and Applied Information Technology, 2014, Vol. 59 No. 1."