ISSN No:-2456-2165

Text Summarization using Word Frequency

Vikas J. Yadav, Tarun M. Pandey, Harsh M. Rathore and Aakash R. Pandey
Department of Computer Engineering
Thakur College of Engineering and Technology
Mumbai, Maharashtra, India

Abstract:- In today's world, where lot of information is accessible on the Internet, it is of utmost importance to provide a better method to retrieve fast and accurate information. We face a lot of difficulty for extracting the brief information about a large document of text. On the internet we have a lot of resources available. So there is a problem of finding appropriate documents from the resources that are available, and extracting useful information from all the available resources.

We can use automatic text summarization to solve the following problem and it can play an important role in summarizing large documents and save time. Text Summarization can be explained as the operation of extracting meaningful collection of sentences which can be replaced in place of large text documents without the meaning being changed. Before starting about the Text summarization, we shall first know what a Summary is, it is a brief text that is produced from a document of text that conveys all the important information in the native text. The main objective of automatic summarization is representing the source data into a brief version that is a summary. By having a summary it greatly reduces the time spent reading the entire document. Text Summarization methods are of two types extractive and abstractive summarization. In extractive summarization methods we select important data from the source data and adding them up to make a summary. The basic idea in Abstractive summarization is to understand the meaning of the phrase or paragraph and rewriting everything in the natural language. Text summarization can be categorized as Inductive and Informative. Inductive summarization represents the main idea from the source text . Inductive summarization generally produces a summary of five to ten sentences. The informative summarization gives correct and brief information about the primary text. Informative summary comprises of only ten to twelve percent of the main text.

I. INTRODUCTION

A. Need

Industry professionals need to search through a pile of documents every day to stay up to date and a large amount of time is spent just to find out what document is important and what is not. By extracting important sentences and creating meaningful brief summaries, it is possible to find out whether a document is worth reading. Students and reviewers are can reap the benefit of text summarization. With the help of text summarization we will be able to generate automated synopsis for our books and research papers in lucid and brief manner while still remaining true to the source material.

B. Related Theory

Synopsis of a huge text is generated using text summarization techniques which is a sub domain of Natural Language Processing(NLP). Text can be summarized using two simple methods, they are NLP based method and deep learning based method. Out of the aforementioned techniques of text summarization we have mentioned an NLP-based technique dependent on word frequency.

The popularity of Automatic Text Summarization rose in the early 1950's. Hans Peter Luhn published a paper in 1958 titled "The automatic creation of literature abstracts". To obtain meaningful sentences from the literature he utilized methods like word and phase frequencies for summarization. In 1969 Harold P Edmundson did a few significant researches in the field of automatic text summarization. To extract influential sentences for text summarization he used features such as the presence of words used in the title appearing in the text, location of the sentences, and cue words. Then after this topic gained lots of attention in the field of Research & Development.

Later developments were greatly influenced by the rise of Machine Learning techniques in 1990's that used different techniques to create literature abstracts. In the beginning of development many systems depended on Naive-Bayes and assumed feature independence while the remaining systems focused on learning algorithms that make no independence assumptions. To refine extractive summarization log-linear models and hidden Markov models are being used. Neural Network and common words in search engine queries are emerging methods to further improvise document summarization.

C. Applications

This model can be used in areas where there is a huge amount of data to be analyzed and a conclusion or an output is to be drawn. Data nowadays is increasing in an exponential rate. This huge volume of data gives rise to a problem of analyzing it and drawing various profitable results from it. This is where summarization comes into picture. Basically summarization is a technique in which large dataset is

summarized into small countable lines of data, so that a user could get something meaningful out of it in a short span of time.

This model also can help in reviewing websites, blogs, news articles, webpage and books. Data nowadays is generated from a lot of source as mentioned above. There are a lot of categories in which data may be segregated but segregating data with this huge volume is not imaginable. Automatic text summarization gives us the power of segregating it in the different categories as per the requirements.

We can use this project to understand a complete book in a short span of time and also develop our review based on that summary. Researchers and scientists need to go through a lot of scientific papers and patents, summarization tool may help them to reduce the time required to skim through the article which in turn will increase the productivity and assist them in new discoveries. Lawyers have large amount of case files and going through all of them is a tedious job. This tool will help in summarizing these case files data and hence the lawyer may be able to understand the case completely in a short span of time.

II. TEXT SUMMARIZATION FEATURES

The key sentences from the reference material is identified and extracted by the text summarizer, this sentences are put together in a sequence to form a meaningful summary. Features as discussed below can be utilized for selection of key sentences in Table 1.

Features	Description
Term Frequency	Salient terms provided by statistics are based on term frequency, thus salient sentences are those words that occur repeatedly. The score of sentences depends on how regularly a word occurs. TF IDF is the most popular method for evaluating word frequency.
Location	The start and end of the paragraph plays an import role in explaining the ideology of the paragraph. So, it is most likely to be a part of the summary.
Cue Method	Effect of positivity or negativity of word on the sentence weight to indicate importance or key idea such as cues: “in summary”, “in conclusion”, “the paper describes”
Title/ Headline word	Words that occur in the title and headline of a document which also occur in sentences are positively related to the topic. Hence, it is important in deriving a summary.
Sentence length	Often very long and very short sentences are not appropriate. Thus To manage the size of the summary it might be excluded.
Similarity	It indicates similarity between the sentence and title of the document, and similarity between the sentence and remaining sentence of the document.
Proper noun	If a document is about a place, person or organization proper nouns become important. So sentences in summaries should have proper nouns for such cases.
Proximity	The gap in the middle of text units where entities occur is a determining element for building association between entities.

Table 1:- Text Summarization Features

III. TEXT SUMMARIZATION TECHNIQUES

Method	Description
Tree based Method	-Representation of the text document is done using dependency tree. -For generating summary it uses multiple techniques such as language generator.
Template based Method	-The whole document is represented using a template. -Patterns are matched to identify bits of text which will then be mapped into template slots.
Rule Based Method	-Representation of the documents is in the form of list of aspects and categories. -All the patterns and rules have to be manually written.

Table 2:- Abstarctive Text Summarization Methods: Structured Based Approaches

Method	Description
Multimodal Semantic Model	-This model captures concepts and relationship among concepts and it is built to represent the concepts of multimodal documents.
Information Item Based Method	-Summary is generated from the abstract representation of source document rather than sentences from the source document.
Semantic Graph Based Method	-A graph known as Rich Semantic Graph (RMG) is generated to summarize a document.

Table 2:- Abstractive Text Summarization Methods: Semantic Based Approaches

Method	Description
Cluster Based Method	-Various themes present in the document should be addressed separately. -If the document is a group of totally dissimilar topics then clustering is necessary in order to create a meaningful summary. -In cluster based method the sentence selection is based on the similarities.
Machine Learning approach	-Summarization process comes under classification problem, sentences are distinguished depending on their characteristics into summary sentences and non-summary sentences. .
Text summarization With Neural Networks	-Neural network is thought on how to gain an understanding on different types of sentences and decide if the sentences are important for the context of summary.

Table 3:- Extractive Text Summarization Techniques

❖ *Text Summarization Steps*

➢ *Transform Paragraphs to Sentences*

Entire paragraph is broken into individual sentences. It can be done by dividing the paragraph at every occurrence of a full stop.

➢ *Text Preprocessing*

Text Preprocessing involves the removal of numbers, stop words and special characters from the sentences that we found in the previous step.

➢ *Tokenizing the Sentences*

All the words in the sentences are tokenized in a numerical manner.

➢ *Find Weighted Frequency of Occurrence*

In the next step we find the weighted frequency of occurrences of all the words. The weighted frequency of a word is given by, frequency of the word divided by the frequency of the word that occurs the most number of times.

➢ *Replace Words by Weighted Frequency in Original Sentences*

The next step is to replace the words with their weighted frequency and find the score for the sentences. It is important to keep in mind that the words which were removed during pre-processing such as stop words, references, etc. have the weighted frequency as zero.

➢ *Sort Sentences in Descending Order of Sum*

The last step of this technique is to find the sentences which contain the words with the most weights i.e. the sentences with the highest scores. After choosing the sentences with high scores they are arranged in the descending order.

IV. CONCLUSION

Due to the requirement for compressed and meaningful synopsis of a topic, because of the colossal amount of information obtainable on the internet, text summarization as a branch of NLP is growing rapidly. Text summarization can be wielded by business analyst, government organizations, teachers, development Researchers, marketing executive and students also. Precise information makes the searching process more accurate and less time consuming. Users require this to process the information in limited time period. We above explain the Extractive Text summarization Technique in detail. This technique can be used in both commercial as well as research community. It is simpler and less time consuming compared to the Abstractive text summarization techniques but the resultant summary is less accurate and meaningful compared to Abstractive methods.

REFERENCES

[1]. Aditi Konge, Manali Sarkar, Rashmi Hatwar, Vrushali Jain “Automatic Recapitulation of Text Document” in the Journal of IRJET published in Volume: 05 Issue: 03 in March-2018
[2]. Deepali K. Gaikwad and C. Namrata Mahender “A Review Paper on Text Summarization” in the journal IJARCCE published in Volume: 05 Issue: 03 in March-2016
[3]. Reeve Lawrence H., Han Hyoil, Nagori Saya V., Yang Jonathan C., Schwimmer Tamara A., Brooks Ari D., “Concept Frequency Distribution in Biomedical Text Summarization”, ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, USA,2006.
[4]. https://stackabuse.com/text-summarization-with-nltk-in-python/
[5]. https://link.springer.com/chapter/10.1007/11846406_37