International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 3
March 2018

Modified Statistical Approach for Summary Generation

Aditi Shashank Chikhalikar
Lecturer, Patkar- Varde College, Goregaon (W)
Mumbai

ABSTRACT —*Due to abundant amount of data/information available, information overload is the problem faced by everyone irrespective of his or her domain. This leads to difficulty in understanding data and making timely decisions. The ability of generating condensed version of available information is need of time. Many times only very small part of large document is indeed very useful. The proposed algorithm of Summary Generation will be helpful in getting the important and useful sentences for the entire document. Thus reducing the time needed to read the document. One can generate summary using proposed algorithm and use the summary generated for deciding whether to read the document or not. Thus, it also helps to decide usefulness of document. There are two categories of summarization linguistic and statistical. Linguistic approach uses knowledge about the language while statistical approach operates by finding the important sentences using statistical methods. The proposed algorithm uses the statistical approach to text summarization. Section I gives introduction to and its need. Section II describes in detail the approach followed for generating summary. The approach uses sentence separator for separating sentences followed by word separator. The stop words (like a, for often etc.) are eliminated before calculating word frequency. The sentences will be ranked based on their scores generated by scoring algorithm. Flexibility of size to summary to be generated is given to the user. Addition of threshold, some heuristic methods based on observations made help to increase the quality of summary generated.*

KEYWORDS— Summary, Information overload, stop words, word frequency, ranking, scoring algorithms.

I. INTRODUCTION

With exponential increase in amount of data and information available, it is important to have ways of filtering and extracting information in nick of time, without being overwhelmed by volume of information.

Generating Summary for these documents can provide a solution. Automatic-Text Summarization is a technique used to generate summaries of electronic documents. It may provide a solution to the information overload problem [1]. Summarization can be used for summing up large documents, multiple documents on same topic with some repeated and some unique content, results from search etc. One can generate summary for deciding whether to read the document or not. Thus it also helps to decide usefulness of document and gives a very condensed overview of the documents. Summarization can help us in many ways, saving our time in day to day work activities. [2,3]

This can either be a generic summary, which gives an overall sense of the document's content, or a query-relevant summary, which presents the content that, is most closely related to the initial search query. [4] Automated document summarization dates back at least to Luhn's work at IBM in the 1950's [5].

A summary can be defined as a text that is produced from one or more texts, that contains a significant portion of theInformation in the original text and that is no longer than half of the original text(s) [6]. According to [7], text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or user) and task (or tasks). A human summarizes the information and write a shorter version with new words, extracting the essence of the original text. However the computer has not been able to match this standard, as it simply generates less redundant version of original text.

A. Approaches

There are two categories of approaches for summarization, text abstraction and text extraction.

Text extraction means to identify the most relevant sentences or phrases in one or more documents, often using standard statistically based information retrieval techniques Shallow natural language processing and heuristics may also be used to enhance the results. These sentences or phrases are then extracted and pasted together to form a non-redundant summary that is shorter than the original. It thus preserves the original

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 3
March 2018

wording and structure of the source text. Sometimes the extracted fragments are post-edited, for example by deleting subordinate clauses or joining incomplete clauses to form complete clauses. Extraction based approach often uses Statistical Methods hence also termed as Statistical Approach. [2,3]
Text abstraction is more challenging task and is quite similar to what humans do when writing an abstract. This approach involves parsing the original text in a deep linguistic way, interpreting the text semantically into a formal representation, finding new more concise concepts describe the text and then generating a new shorter text, an abstract, with the same information content.[3]
This paper has elaborated on Extraction based approach/ Statistical Approach.

II. METHODOLOGY

This section illustrates the proposed algorithm for summary generation.

A. Text Pre-processing

The first step to generate good summary is pre-processing. The pre-processed text is considered for further analysis. This stage will include: Stop-word elimination, Tokenization and stemming.

1. Stop-word Elimination.

In pre-processing of a text, an intermediate representation of text is obtained. One of the pre-processing stages consists in eliminating stop-words or empty words from the text. There is a set of empty words in every language, common to all domains which are easily identified, for example, articles, prepositions, conjunctions, etc. Although they can be verbs, adjectives and adverbs the words that are too frequent in the documents in a particular collection are not good discriminators. In fact, it is considered that a word that appears in at least 80% of the documents of a particular collection is useless for purpose of retrieval. [11] These words are considered empty and normally are removed to avoid being considered as potential. The aim is to reduce the content of the text to more specific expressions, containing only the words that are useful and meaningful for the generation of automatic summaries. For algorithm stated in this paper 635 stop words have been decided.

2. Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens will be input for further processing. [12]
In the system proposed in this paper, first text is broken into sentences. Simple rules like sentence ending are determined by a dot (.) or Exclamation (!) or Question mark (?) etc. can be used to separate sentences. While words are separated using delimiters like space (' ', ',', ';', ':', '"', '(', ')', '[', ']', '{', '}', '.', '!', '?', '/', '•').

3. Stemming

Stemming technique consists in obtaining the root of words, so that the text processing is conducted on the roots and not on the original words. This technique allows relating of more terms in the document. Assumption is that the words having the same root represent the same concept. Basically, the process of stemming of the words is realized for reducing to a minimum common portion of a word called stem. The stem is the portion of the word which is left when after the removal of its affixes, prefixes and suffixes. Once implemented stemming, the document will contain only the roots of the words. The first stemming algorithm was developed for the English language, and then was adapted for the Spanish language. The algorithm Porter is the most commonly used for the English language. In general, these algorithms are based on a simple set of rules that cut off words to obtain a common root. [11,13] Algorithm in this paper also uses Porters Algorithm for stemming.

B. Scoring and Ranking

This step involves calculating the word frequency, and then assigning the scores to each sentence followed by ranking the sentences according to the scores.

1. Word-Frequency Calculator.

This calculates the number of times a word appears in the document, stop-words will not be taken into account as they have been eliminated. The number of sentences that word appears in the document will be considered. As some sentences may have more than one occurrence of the word.

2. Scoring (Sentence Scorer)

This step involves assigning the scores to each sentence followed by ranking the sentences according to the scores. Scoring algorithm determines the score of each sentence. This will determine the score of sentence

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 3
March 2018

based on the number of distinct words present in the sentence. The number of distinct words present in the sentence is termed as WordCount of sentence. The frequency calculated in above step is taken into account. But if the frequency of a word is more than threshold value then score for such words will be reduced; considering that they are quite obvious and hence occurring in high frequency. Threshold value can be calculated as average of a 70 percent of top frequency word.

threshold = (cntWords.Max() * 70) / 100;

(Max of cntWords will give you maximum time a word occurring in given text.)

If the word in sentence is one of the keyword entered by user then the score of sentence will be increased by threshold value calculated above. The reason is to give high weightage to sentences having keywords. The additional weightage is considered as threshold value so that weightage will be assigned according to document under processing rather than statically assigning the value. Statically assigned value may not give consistent results in all cases. The score generated from these frequency related considerations will be termed as FreqScore (in this algorithm).

3. Ranking (Sentence Ranker)

Then the sentences will be ranked according to the scores calculated. It is observed during user inaction and analysis of data collected during study by author; that generally irrespective of the score human users tend to select 1st or 2nd sentence of the document. To enhance the summary this heuristic approach is also included in this algorithm. Here first two sentences are compared and one with highest score is considered as a part of summary. Now the remaining sentences are selected based on the rank to generate rest of the summary.

C. Summary Generation

In this last stage, the summary is generated by picking up the required number of highest scored sentences. Length of summary may be given by user or it will be considered as 1/3rd of file length. The generated summary are displayed to user on screen and saved in a file.

D. Evaluating the Algorithm [3]

There are two properties of the summary that can be measured to evaluate summary generation systems they are the Compression Ratio, and the Information Retention Ratio (IIR). The compression ratio gives the amount of compression achieved. Whereas IIR will help to infer the quality of summary on the basis of retention of important words in summary wrt.text. IRR is also referred to as omission ration. The equations can be given as follows:

Compression Ratio: CR = (Length of Summary) / (Length of Text)

Information Retention Ratio: IRR = (Total No. of occurrences' of all distinct words in Summary) / (Total No. of occurrences of all distinct words in Text)

III. APPICATIONS

Among various applications of application areas for automatic text summarization, the most common one is information retrieval. Information retrieval systems rank and present documents based on measuring relevance to the user query. Not all documents retrieved by the system are likely to be of interest to the user. Presenting the user with summaries of the matching documents can help the user identify which documents are most relevant to the user's needs. [4]

Many a time's information on same topic is available simultaneously on many media channels in different versions. Customization of information for different channels and formats is an immense editing job that notably involves shortening of original texts. Automatic text summarization can automate this work completely or at least assist in the process by producing a draft summary. [2,3]

In Hirst et al. (1997) a system that from medical digital libraries produces user-adapted information towards individual patients' specific needs, summarized from information on surgery of breast cancer to living with diabetes but also general health education is presented.

Another system is PERSIVAL, which is described in McKeown et al. (2001). PERSIVAL generates user-adapted information both for patients and physicians, and uses as input the patient record of the patient to find what topics the generated text should contain. PERSIVAL then searches for the relevant information in external resources and summarizes it to the relevant level of the user. The text that is constructed for patients'

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 3
March 2018

origins from several consumer health texts, while the text constructed for physicians is collected from medical journal articles. [2]

Some of current uses of summarization [10-8]:

1) Multimedia News Summarization: This involves summarizing data from different sources.

2) Producing Intelligence Reports: Given a wide range of documents, an intelligence analyst may wish to read a biography of a person. A system exists which creates a dossier of information on a person from a text collection.

3) Text for Hand-held devices: Due to the limited size of displays on WAP phones and palm-top computers, it is useful to condense text found in web pages browsed.

4) Convenient Text-to-Speech for Blind people: The idea here is to scan in a page from a book, and then read out a summary of the page rather than the entire text.

5) Summarizing Meetings: Combining summarization with automatic speech recognition produces a system which summarizes the salient points of a meeting.

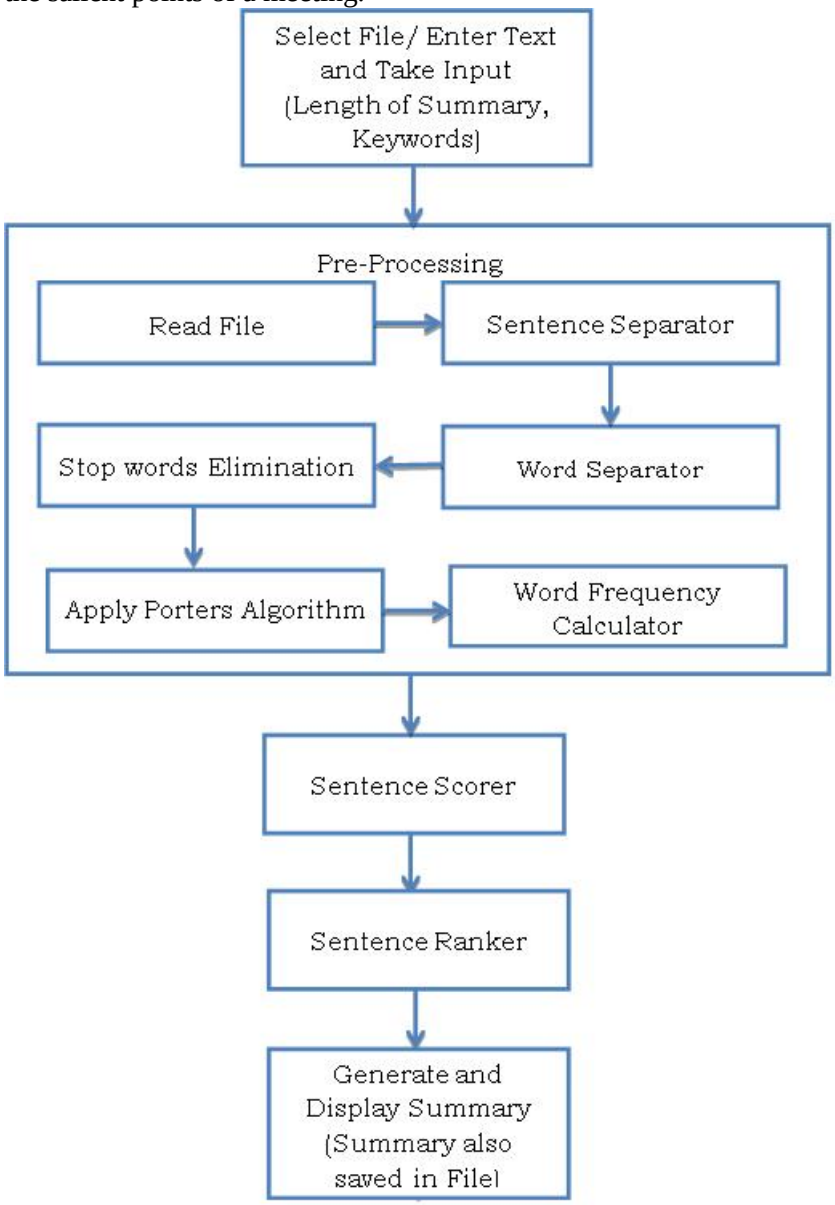

Figure 1: Algorithm for Modified Statistical Approach for Summary Generation

IV. ADVANTAGES AND DISADVANTAGES

Extraction based approach has both advantages and disadvantages. They are easily adapted to large documents as they are limited to the extraction of sentences however, the resulting summaries may be incoherent. Abstraction approaches, on the other hand, provide more sophisticated summaries, which often contain material that enriches the source content such as those needed for wireless personal digital assistants (PDAs) and similar technologies [1]. Extraction based approach may not provide best results here. However, the strengths of this approach are that it's robust, and provides good results for query based summaries. In addition, it overcomes the disadvantages like slow speed; need to scale up to robust open-domain summarization of abstraction approaches. Major disadvantage is its inability to manipulate information at abstract levels. [9]

V. CONCLUSIONS

Automating Summary generation tool is the need of hour owing to ever increasing amount of information and the requirement to determine important information in short time period. Abstraction approaches provides more sophisticated summaries as compared to Extraction approaches. On other hand it overcomes the disadvantage that Abstraction approach is domain specific to a certain extent. Thus if we want to develop an portable, open-domain summery generation technique we can for Extraction based approach. Algorithm stated in this paper aim to enhance the efficiency of Extraction approaches by addition of threshold- making summary text specific. Addition of heuristics methodsmakes the summary more closer to that generated using Abstraction approaches.

REFERENCES

[1]. Udo Hahn , Inderjeet Mani "The Challenges of Automatic Summarization" Albert Ludwigs University Mitre Corp, IEEE Computer, Vol. 33, No. 11, pp. 29–36, 2000.
[2]. Martin Hassel and Hercules Dalianis "Portable Text Summarization", Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden.
[3]. MARTIN HASSEL "Evaluation of Automatic Text Summarization", Licentiate Thesis Stockholm, Sweden 2004.
[4]. Jade Goldsteiny, Mark Kantrowitz, Vibhu Mittal, Jaime Carbonelly "Summarizing Text Documents:Sentence Selection and Evaluation Metrics", Copyright 1999 ACM 1-58113-096-1/99/0007
[5]. Luhn, P. H. "Automatic creation of literature abstracts", IBM Journal (1958), 159,165.
[6]. Hovy, E. H. "Automated Text Summarization". In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583-598.Oxford University Press, 2005.
[7]. Mani, I., House, D., Klein, G., et al ."The TIPSTER SUMMAC Text Sum-marization Evaluation". In Proceedings of EACL, 1999.
[8]. ParthaLal "Text Summarization", June 13, 2002
[9]. Stephan Busemann, DFKI GmbH "Automated Text Summarization", Language Technology at ACL/COLING1998 I, WS 2007/2008
[10]. Inderjeet Mani "Automatic Summarization. Natural Language Processing." John Benjamins Publishing Company, 2001.
[11]. YuliaNikolaevnaLedeneva "Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization" Thesis INSTITUTO POLITECNICO NACIONAL (National Polytechnic Institute, IPN) Mexico.
[12]. http://tint.fbk.eu/tokenization.html
[13]. http://snowball.tartarus.org/algorithms/porter/stemmer.html