2013 Third International Conference on Advances in Computing and Communications

A Novel Technique for Efficient Text Document Summarization as a Service

Anusha Bagalkotkar
Dept of Information Technology
NITK Surathkal India
anusha.bagalkotkar@gmail.com

Ashesh Khandelwal
Dept of Information Technology
NITK Surathkal, India
ashesh.k10@gmail.com

Shivam Pandey
Dept of Information Technology
NITK Surathkal, India
shivam.pandey1991@gmail.com

Sowmya Kamath S
Dept of Information Technology
NITK Surathkal, India
sowmyakamath@ieee.org

***Abstract*—Due to an exponential growth in the generation of web data, the need for tools and mechanisms for automatic summarization of Web documents has become very critical. Web data can be accessed from multiple sources, for e.g. on different Web pages, which makes searching for relevant pieces of information a difficult task. Therefore, an automatic summarizer is vital towards reducing human effort. Text summarization is an important activity in the analysis of a high volume text documents and is currently a major research topic in Natural Language Processing. It is the process of generation of the summary of an input document by extracting the representative sentences from it. In this paper, we present a novel technique for generating the summarization of domain-specific text from a single Web document by using statistical NLP techniques on the text in a reference corpus and on the web document. The summarizer proposed generates a summary based on the calculated Sentence Weight (SW), the rank of a sentence in the document's content, the number of terms and the number of words in a sentence, and using term frequency in the input corpus.**

Keywords-Text Summarization; POS Tagging, Knowledge Extraction, Natural Language Processing

I. INTRODUCTION

Automatic text summarization has drawn substantial interest from researchers and software developers since it provides a solution to the information overload problem for users in this digital era of the World Wide Web. Readers are overloaded with too many lengthy text documents when they are more interested in shorter versions.

Two fundamental techniques are identified to automatically summarize texts, i.e. abstractive and extractive summarization [1]. Complex summarization techniques are generally based on abstraction. It uses computer generated analyses and synthesis of the source documents into a completely new documents. Not only is the summarized document shorter but also cohesive, readable and intelligible. Multidisciplinary approaches in information retrieval, linguistics, machine learning and artificial intelligence have been applied to achieve the abstractive summarization.

In contrast to abstraction, which requires using complex techniques from natural language processing (NLP), including grammars and lexicons for parsing and generation, extraction can be easily viewed as the process of selecting important excerpts (sentences, paragraphs, etc.) from the original document and concatenating them into a shorter form. The emergence of WWW applications on one hand and the exponential increase of the internet documents' sizes on the other hand are increasingly making searching for useful information a very difficult task. This raises interest for automatic document summarization [2], which helps creating a concise summary of the document(s) to the user. For instance, an automatic text summarizer can provide a briefer version of a document to aid the user to quickly determine whether such document is relevant to him or not. However, despite the paramount importance of such systems, the performance of developed systems is very limited due to the various challenges encountered in information processing [2]. In spite of the increasing advancement in natural language processing tools that help in tokenizing sentences, extracting desired entities and quantifying the relationship among various terms or sentences, a generic design of an automatic summarizer is still very challenging.

Text summarization techniques can also be classified on the basis of volume of text documents available in the text database. If summarization is performed for a single text document then it is called as the single document text summarization. If the summary is to be created for multiple text documents then it is called as the multi document text summarization technique. [3]

The organization of the paper is as follows. Section II presents some background to the work presented and section III discusses related work. In Section IV, we discuss the proposed approach of text summarization and experimental results in Section V followed by conclusion and future work in Section VI, and references.

II. BACKGROUND

Currently, the number of documents retrieved by Web Search Engines is already beyond the capacity of human analysis due to the fact that hundreds of pages of search results are generated for most input queries. Thus document retrieval is not sufficient and we need a second level of abstraction to reduce this huge amount of data - the ability of summarization.

978-0-7695-5033-6/13 $26.00 © 2013 IEEE
DOI 10.1109/ICACC.2013.17

Automatic text summarization condenses text contents into most important concepts and ideas under a particular context. This technology may be helpful to identify topics, categorize contents, and summarize documents. However, most previous work on automatic text summarization has emphasized on information abstraction and extraction. Some well-known approaches, like TF/IDF (Term Frequency/Inverse Document Frequency) [5], which summarizes a text based on term frequency weight that is assigned to each term, neural network system for text summarization, statistical models, and so on, usually rank sentences and select sentences with higher ranking Score as the summary.

There are two properties of the summary that must be measured while evaluating summaries and summarization systems – the Compression Ratio, which is a measure of the length of the summary when compared to the original, and the Retention Ratio or Omission Ratio, which is a measure of how much of the document's central information is retained in the summary. [6]

Semantic similarity [11] is a concept frequently employed in determining the ranking of a term or sentence. A set of documents or terms within term lists are assigned a metric based on the likeness of their meaning / semantic content. Various semantic similarity techniques are available which can be used for measuring the semantic similarity between text documents. Semantic similarity methods are classified into four main categories, Edge Counting Methods that measure the similarity between two terms (concepts) as a function of the length of the path linking the terms and on the position of the terms in the taxonomy, Information Content Methods to measure the difference in information content of two terms as a function of their probability of occurrence in a corpus, Feature based Methods to measure similarity between two terms as a function of their properties (e.g., their definitions) or based on their relationships to other similar terms in the taxonomy and Hybrid methods that combine the above three mentioned methods for calculating the semantic similarity.

III. RELATED WORK

There is a lot of effort in the field of achieving effective text summarization. Nagwani et al. [6] proposed a frequent term based text summarization algorithm that first processes the document to be summarized by eliminating stop words and by applying stemmers. Next, term-frequent data is calculated from the document and frequent terms are selected, and for these selected words the semantic equivalent terms are also generated. Finally, all sentences in the document that contain the frequent terms identified and their semantic equivalents are filtered for summarization.

Guangbing et al. [7] introduced a personalized text-based content summarizer to help mobile users to retrieve and process information more quickly, as per their interests and preferences. It is based on probabilistic language modeling techniques adapted to build a user model and an extractive text summarization system to generate a personalized and automatic summary for mobile learning.

Aksoy et al. [8] proposed an idea of using Semantic Role Labeling (SRL) on generic Multi-Document Summarization (MDS). Sentences are scored according to frequent semantic phrases and the summary is formed using the top-scored sentences. This method used a term-based sentence scoring approach to investigate the effects of using semantic units instead of single words for sentence scoring. Then scoring metric is integrated as an auxiliary feature with the intention of examining its effects on the performance.

Rushdi et al [9] put forth a novel technique for summarization of domain-specific text from a single web document that uses statistical and linguistic analysis on the text in a reference corpus and the web document is presented. The proposed summarizer used the combinational function of Sentence Weight and Subject Weight to determine the rank of a sentence. It used the number of terms and number of words in a sentence, and term frequency in the corpus for summarization and about 30% of the ranked sentences were considered to be the summary of the web document. Three web document summaries using the proposed technique were generated and compared with the summaries developed manually from 16 different human subjects.

Foong et al. [10] developed a hybrid Harmony Particle Swarm Optimization (PSO) framework for an Extractive Text Summarizer to overcome high processing load. Their objective was to find out if the proposed PSO model was capable of condensing original electronic documents into shorter summarized texts more efficiently and accurately than the alternative models. Their empirical results showed that the proposed hybrid PSO model improved the efficiency and accuracy of composing summarized text.

IV. PROPOSED SYSTEM

The proposed method can be described in four distinct phases as shown in Figure 1.

a) Generation of the list of frequent words: In the first step the document which is required to be summarized is given by the user and then processed by eliminating the stop word. After eliminating stop words the term-frequent data is calculated from the document and frequent terms are selected which are used to generate document summary.

b) Sentence Generation: After the selection of frequent terms, the sentences which contains the frequent terms are extracted from the given document. For generating a summary of document the compression ratio is set to 1/3 of the original text i.e., for every three sentences in text one sentence is introduced as summary.

c) Update details in database: When the summary is generated then its details is stored in the database and is available to the user for information analysis.

d) Setup Web Service: A web service to provide summary of given text will be set up. The Web Service client will send request message consisting of document then the server sends the summary as the response message.

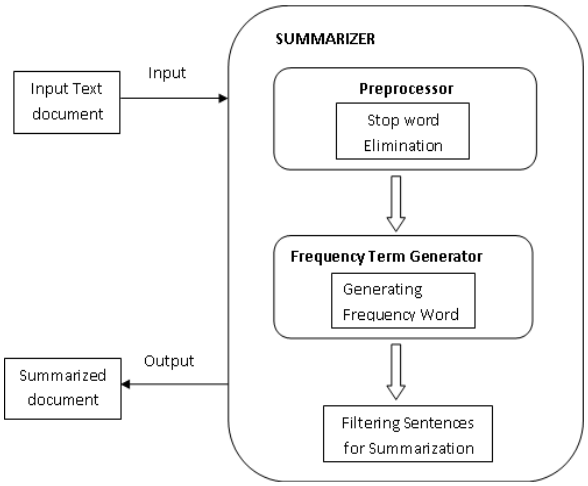

Figure 1. The Proposed System

The overall proposed system is shown in Figure 1, where all the steps are depicted in a sequential manner. The system is divided into three major parts, an input module for the text document to be summarized, a summarizer algorithm and the output module. The summarizer algorithm is further divided into the three parts – the text pre-processing module, frequent terms generation module along with the semantically similar terms and sentence filtering module for summarization.

Algo SingleDocSummarizer
Input:
1. Text Data for which Summary is required.
2. Value of N – for generating top N frequent Terms.

Output:
1. Summary for the Original Text Data.
2. Compression Ratio.
3. Retention Ratio.

Steps:
1. Data Preprocessing Phase
- *Retrieve data*
- *Eliminate Stop Word*

2. For the entire text content
- *Get the N frequent Terms*
- *Generate Term-Frequency List*

3. For all N-Frequent Terms
- *Generate Sentences from the Original Data*
- *If the sentence consists of a term that is present in frequent-terms-list then*
  - *add sentence to summary-sentence-list.*

4. Calculate Compression Ratio and Retention Ratio

Figure 2. Single Document Summarizer Algorithm

The text summarization technique which is implemented in the proposed system is as follows. In the preprocessing stage, the unstructured text is converted into structured text. The stop words are then removed and the document is parsed to collect all the words along with their frequencies. The second stage is to extract the important key-phrases in the text by implementing a new algorithm through which extracting high frequency words. The system uses the extracted keywords to select the important sentence. The third stage deals with extracting the sentences with highest rank from the input text.

The semantic similarity based on single document summarization is shown in Figure 2. The algorithm takes two input parameters – the input text document and number of frequent terms. As the output it generates a summarized text document along with the two measures compression ratio and retention ratio, which is further explained in the next section. Single Document Text Summarization is generated using Frequent Terms and Semantic Similarity.

V. EXPERIMENTAL RESULTS

5.1. Experimental Setup

The algorithm is the functionality provided by a Web service which can be invoked on document submission. The user entered text or uploaded document is sent to the server via a SOAP request, where the service is invoked. After the text is accepted, the Java Web service client provides the summary through a SOAP Response. Fig. 3 shows the Summarizer Web Service in action. The input text is embedded within a SOAP Request and resultant summary is sent back by embedding it within the SOAP Response. After accepting the request the web service provides a SOAP Response consisting of the summary for the text which has been provided by user.

5.2. Evaluation Parameters

There are two properties of the summary that must be measured while evaluating summaries and summarization systems – the Compression Ratio, i.e. how much shorter the summary is than the original. The compression ratio (CR) can be calculated using equation (1).

$$CR = (\text{length of summary S}) / (\text{length of input text T}) \quad (1)$$

5.3. Experimental Results

Figure 3 shows a snapshot of the Summarizer Web Service being invoked by user through the service client and the summary being returned by the service along with the number of words in the sample text, summarized text and the compression ratio. In the example shown, the input sample had about 180 words, while the summary contained 68 words. Hence the compression ratio was calculated to be about 33%. Alternatively, we can allow the users to specify the compression ratio as per their requirement or comfort based on which the summary can vary in length.

In order to evaluate the service further, documents and texts of various lengths were given as input to the Web service to generate the summary. The purpose of this was to calculate the compression ratio achieved by the algorithm implemented by the Web service for summary generation. Table 1 shows the results of the conducted experiments. In general, the compression ratio was observed to be 1/3 i.e. the

summarized output was about 33% lesser in length than the input text.

TABLE 1. COMPRESSION RATIO EVALUATION

Document length (in lines)	Summary length (in lines)	Time taken (in ms)
6	2	62
11	3	109
31	10	171
55	18	358
100	33	603
177	68	1097
250	82	1329

VI. CONCLUSION AND FUTURE WORK

Since there is vast amount of textual information available on Web which cannot be analyzed by humans, a service oriented approach can be useful for retrieving important information from text. In this paper, we developed a single document frequent terms based text summarization algorithm for summarizing the given text by extracting relevant sentences for creating the summary. The proposed method is basically an extraction based approach.

Future work includes applying the same concept for multi-document texts. After implementing it for multi-document texts the Web service will be able to provide a summary of multiple documents by extracting frequent sentences calculated based on frequent terms that are present in multiple documents, thus providing a summary to the user for similar text documents. The concept of semantic similarity to extract any words which are semantically equivalent to the frequent terms can be applied, thus the sentences with both frequent and semantically equivalent terms can also be incorporated while generating the summary.

REFERENCES

[1] Nenkova, Ani, and Kathleen McKeown. Automatic summarization. Now Publishers Inc, 2011.

[2] Mani, Inderjeet, and Mark T. Maybury. Advances in automatic text summarization. the MIT Press, 1999.

[3] Goldstein, Jade, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. "Multi-document summarization by sentence extraction." In Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4, pp. 40-48.

[4] Lal, Partha. "Text Summarization." (2002).

[5] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management (1988)

[6] Kumar Nagwani, Naresh, and Shrish Verma. "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm." International Journal of Computer Applications (2011)

[7] Yang, Guangbing, Wen, Nian-Shing, and Sutinen. "Personalized Text Content Summarizer for Mobile Learning: An Automatic Text Summarization System with Relevance Based Language Model." InTechnology for Education (T4E), 2012 IEEE Fourth International Conference on, pp. 90-97. IEEE, 2012.

[8] Aksoy, Bugdayci, Gur, Uysal, and Can. "Semantic argument frequency-based multi-document summarization." In Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on, pp. 460-464. IEEE, 2009.

[9] Shams, Rushdi, M. M. A. Hashem, Suraiya Rumana Akter, and Monika Gope. "Corpus-based web document summarization using statistical and linguistic approach." In Computer and Communication Engineering (ICCCE), 2010 International Conference on, IEEE, 2010.

[10] Foong, Oi-Mean, and Alan Oxley. "A hybrid PSO model in Extractive Text Summarizer." In Computers & Informatics (ISCI), 2011 IEEE Symposium on, pp. 130-134. IEEE, 2011.

[11] Resnik, Philip. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." arXiv preprint arXiv:1105.5444 (2011)

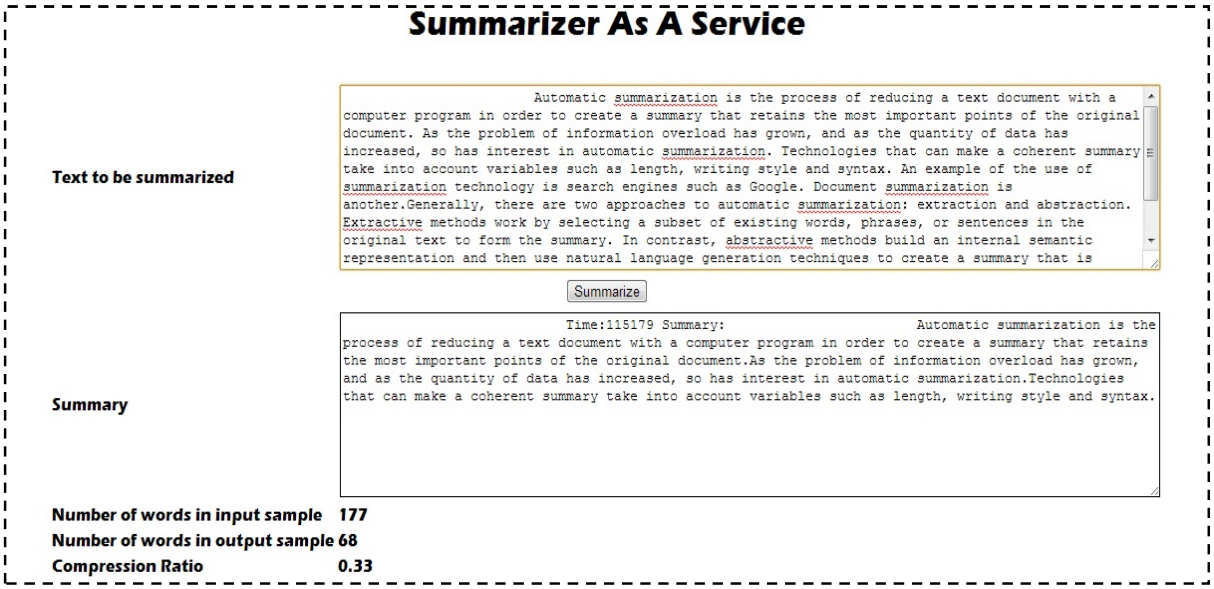

Fig 3: A screenshot of the Summarizer Web Service in action