Proceedings of the 11th INDIACom; INDIACom-2017; IEEE Conference ID: 40353
2017 4th International Conference on "Computing for Sustainable Global Development", 01st - 03rd March, 2017
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

A Study on various Techniques for Automatic Text Summarization

Ayushi Negi
Undergraduate student,
BVCOE, New Delhi, India
ayushin78@gmail.com

Mehak Jain
Undergraduate student,
BVCOE, New Delhi, India
mehak_jain13@yahoo.in

Shilpa Gupta
Assistant Professor, CSE department
BVCOE, New Delhi, India
shilpa.gupta@bharatividyapeeth.edu

Abstract – *With the advent of modern technology and high-speed internet, access to various texts and documents has been highly simplified. But it has given rise to a new problem. Due to the availability of a large amount of data, it has become very difficult to analyse and understand all the information in a given period of time. The growing need of a technology that enables gathering a large amount of relevant data according to the user's requirement and queries in a short period of time has given rise to Automatic Text Summarization. It is an upcoming technology in which a brief summary of a text is generated by the machine without any human intervention. Spanning different techniques in extraction and abstraction, Automatic Text Summarization encompasses various domains. The paper contains a brief introduction of different types of Automatic Text Summarization, followed by a survey of different techniques for it in different domains. The paper discusses the key features of these techniques followed by a comparison of techniques in different domains.*

Keywords – Text Summarization, Extraction, Abstraction

I. INTRODUCTION

The goal of any summarization technique is to generate a summary or abstract from a text. The text may be a single document or spanning over multiple documents [1]. A summary can be defined as a condensed version of the original text, which contains all the relevant points of the document it is summarising organised in a logical manner. Thus a summary must contain all the relevant points mentioned in the text and be shorter in size. The process of generating a summary from the original text is known as summarization.

Automatic text summarization is the art of abstracting key content from one or more information sources using a machine without any human intervention. [2] Hence, it encompasses techniques which teach a machine how to summarize a text into a meaningful summary without taking any human input.

Broadly, we can say that the main goals of automatic text summarization are as follows:

1. Minimization: The main purpose behind the summarization process is that it should save of time and human energy spent in scanning entire documents. This is only possible if the summary is smaller in size as compared to the entire text.

2. Information Integrity: On the other hand, many techniques often render the resulting abstract meaningless as the summary fails to capture the idea behind the text or messes up the information. Hence, maintaining the integrity of the information becomes a prime focus in Automatic Text Summarization.

The summarization process has mainly three phases: The first phase involves extracting the main idea behind the text by analysing the text document. It is followed by the second phase in which the most information rich sentences are selected closely followed by the third phase involving generation of grammatically correct sentences from the output of the second phase.

It can be divided into two types depending on the method of summarization used:

1. Extraction: In Extraction, the machine first uses some algorithm to find the most relevant words or sentences or phrases from a document. These units are then collected and a summary is produced

simply by putting them together in a coherent manner.

2. Abstraction: In Abstraction, the entire text is analysed and the main idea of the text is presented in the summary in a concise and clear manner by using our own words or sentences.

This paper has been divided into sections. Section I is the current section containing the introduction. Section II and III deals with Extraction and Abstraction respectively. Section IV contains the related work reviewing the various techniques proposed. Section V provides a comparative analysis of all the different techniques discussed. Section VI deals with the conclusion of the paper.

II. EXTRACTION

Extraction involves an analysis of the original/source text to gather all the important and relevant lexical units, which may be words or sentences, on the basis of the significance factor assigned to them. The lexical units having significance factor higher than a threshold value are concatenated together to create the summary. The process of Extraction is explained in the following figure (Figure 1).

As shown in Figure 1, an extractive summarizer takes the source text and the various factors used to determine the significant factor as input and outputs a list of sentences ordered on the basis of their significant factor, in a decreasing order. A threshold value is set, and all the sentences having value greater than the threshold are used in the abstract.

Different techniques use different parameters to calculate the significance factor. Luhn HP [3] was the first to use significance factor to determine the importance of a sentence in the text. He used the frequency of word occurrences and the relative position of words (that have a given value of significance) in the sentence as parameters to determine the significance factor of a sentence.

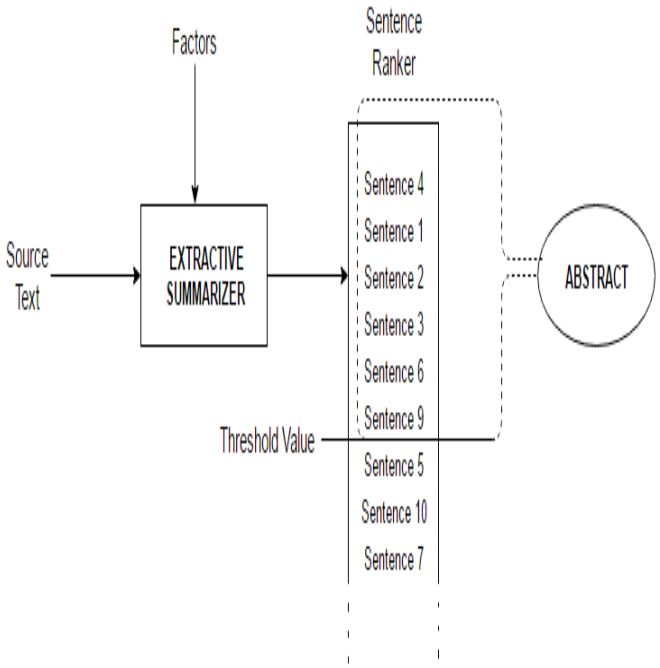

FIGURE 1: EXTRACTIVE SUMMARIZATION

Edmundson HP [4] used a variety of different parameters to calculate the significance factor of a sentence. He used the frequency of occurrence of a word, cue words, the presence of the word in title and location of the sentence to calculate the significance factor.

Kupiec J [5] et al considered the following factors for calculating sentence significance: length of sentence, fixed phrase feature, the position of sentence in a paragraph, word frequency, and uppercase word feature.

Linear combination of these factors is used to calculate the significant factor of a sentence. These sentences are then ranked accordingly and sentences having significant factor greater than the threshold value are used to generate the abstract.

Sentence scoring is one of the most significant task in extractive summarization. Rafael Ferreira [6] reviewed all the major techniques for sentence scoring and tested them on different data sets and found that word frequency, Tf/Idf [7], lexical similarity and sentence length provided the most satisfying results of all the heuristic rules applied for sentence scoring.

The main problem with extraction lies in the fact that it views the text as a set of sentences and words instead of words defining an idea or theme. Many a time, summaries generated from extraction fail to capture the true spirit of the text, missing the idea behind the textual material.

III. ABSTRACTION

Abstraction is another type of text summarization in which instead of parsing sentences and words, the technique attempts to understand the conceptual meaning of the text and write it in own words. Hence, a summary generated by abstraction does not use the sentences used in the document, instead, it understands the document and writes a summary by creating meaningful sentences.

Abstraction techniques can be divided into two broad categories: [8]

1. Structured Based Approach: It uses cognitive schemes to encode the most important information from the text or documents.

2. Semantic Based Approach: This method basically identifies noun phrases and verbs by processing the linguistic data.

Some of the techniques used in abstractive summarization have been discussed in this section.

Structured Based Approach

Barzilay R [9] proposed an approach based on the belief that when repeated information about an event or an occurrence is encountered in multiple documents, it becomes a good indicator that the information is important and thus must be present in the summary. They first identified the similarities between phrases that report the same information. This is done by first creating a dependency based representation, DSYNT. Then all sentence trees which are rooted at verbs are traversed recursively. When two nodes are found to be the same, they get added to the output tree and their children are compared. When an entire phrase is found, it is added to the theme intersection. Additional information such as temporal ordering is also taken into account. These phrases then input into a sentence generator, called FUF/SURGE to generate sentences which are further utilised to create the summary. The usage of a sentence generator and additional information like temporal ordering renders the summary informative and useful. However, the context from the text id not included in the abstract which may prove to be essential in some situations. Also, lack of a proper evaluation method may prove to be problematic in future

Genest PE et al [10] proposed a full abstraction technique relying on Information Extraction, statistical content selection, and Natural Language Generation to accomplish guided summarization. In this multi-document summarization technique, firstly a category is allocated to a ten document group which are required to be summarized. Then this cluster of documents is pre-processed. The pre-processing includes various tasks like normalization of the input text, segmentation of the sentences, etc. The pre-processing task also includes a date resolution engine. This is followed by the generation of abstraction schemes. Each abstraction scheme contains IE rules where IE stands for Information Extraction, content selection, etc. and pertains to a particular theme. These IE rules then identify a large number of candidates for each aspect. The most appropriate amongst these, mostly based on their degree of occurrence, are selected and are used to generate a summary. Finally, a sentence generator takes these phrases and words in their root form as an input and outputs the sentences which constitute the summary. The main advantage of this technique is than it produces a much smaller summary thus increasing the information density in a summary and allowing access to larger amount of data in a smaller span of time.

Semantic Based Approach

Greenbacker CF [11] presented a structure for Abstractive Summarization of Multimodal documents. The suggested framework consists of three steps. In the first step, a semantic model is developed by analysing the text with a parser with the help of knowledge representation focused on ontology. The second step consists of rating of the content on the basis of a metric, Information Density. This is done to identify the most salient and important concepts from the entire cluster of data. In the third and final step, these salient concepts are captured in a semantic model and then expressed as sentences. These sentences form the abstract. The main advantage of the techniques is that it is applicable not only to text but also other media like images.

Moawad IF [12] proposed a graph based approach which follows three phrases in summary generation. The first phase, Rich Semantic Graph Creation Phase, focuses on semantic representation the text as RSG, where RSG stands for Rich Based Graph. It is followed by the RSG Reduction Phase. In this phase, different rules are used to reduce the RSG thus creating an RSG having higher information density. The third phase, the Text Generation Phase, contains the generation of the final summary. In this phase, the semantic representation thus generated is used to create the summary by processing it with the information rich domain ontology. Since this process removes the redundancy in text, thus the information density of the abstract is higher than in most techniques.

IV. RELATED WORK

Various studies and researches are done in the field of text summarization in different domains. Some of the extractive summarization techniques have been discussed in the following section.

Centroid Based Summarization

It is an extractive technique developed for multi-document cluster. The central feature lies in identification of most central sentences in a multi-document cluster. These sentences capture sufficient data connected to the main idea of the documents. In this technique, words having a high

value of TF*IDF [6] (above a pre-defined threshold) are collected in a pseudo-document called a Centroid. The sentences that contain a significant number of words from the centroid are selected and can be further furnished to create the summary. [13]

Radev DR [14] proposed a centroid based summarization system, MEAD. The system accepts the source texts and compression rate as input then uses three features to rank the sentences according to their salience. The sentences are re-ranked based on the redundancy between the sentences. Sentences which are similar to other sentences already in the abstract are disregarded. The above mentioned features are Centroid Score, Position Score and Overlap with first Score. The system has been depicted in Figure 2.

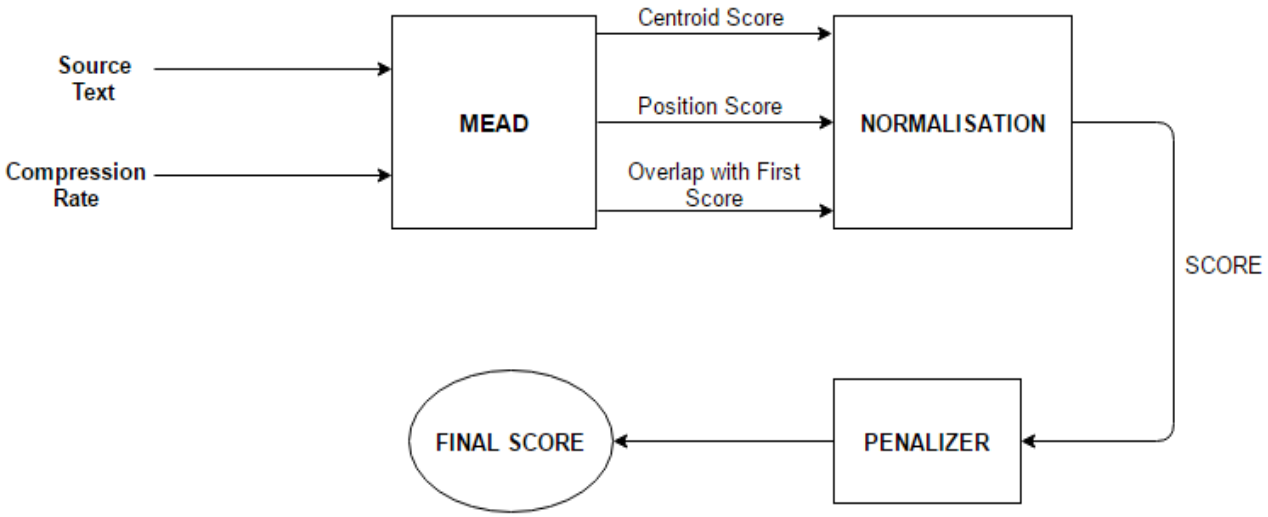

FIGURE 2: A CENTROID BASED SUMMARIZER – MEAD

Centrality Based Summarization

Erkan G [13] proposed a centrality based sentence salience summarization which focuses on the idea that the sentences which are similar to many other sentences are more related to the main theme of the text. To determine similarity, they used a bag-of-words model for the representation of each sentence. For every word in a sentence, the related dimension found in the vector representation of that particular sentence in set equal to the product of the number of occurrences in the sentence and the idf of the word. Following this, the cosine of two the two sentences is then calculated to determine the similarity between the two sentences.

$$idf\ modified\ cosine(x,y) = \frac{\sum_{w \in x,y} tf_{w,x}\, tf_{w,y}\, (idf_w)^2}{\sqrt{\sum_{x_i \in x} \left(tf_{x_i,x}\, idf_{x_i}\right)^2} \times \sqrt{\sum_{y_i \in y} \left(tf_{y_i,y}\, idf_{y_i}\right)^2}}$$

Using these values, a cosine similarity matrix is computed. The significant similarities in the matrix are then utilised to generate a weighted graph in which each edge represents the similarity between two sentences and the nodes represent the sentences. The degree of a node in the similarity graph is then named as the degree centrality of the corresponding sentence. The sentences are then ranked on the basis of their degree centrality and those having value greater than a threshold are selected for the abstract.

Centrality Based on LexRank

Erkan G [13] proposed another approach for finding the centrality, by the name LexRank. There may arise a situation in which an unrelated document may contain some sentences which have high value of degree centrality. For example, if we are applying the graph based technique discussed above on a cluster of physics based documents and the cluster contains a document entirely devoted to medicine. Since the sentences in the medicine document will be related to each other in the graph, these sentences will have a high degree centrality. However, as these sentences are completely unrelated to the topic of the abstract, they should not occur in the abstract. Hence, the centrality based approach faces this hurdle in multi-document summarization. This can be overcome by considering the degree centrality of not only the sentence under consideration but also of all the nodes contributing to the degree centrality of the sentence, as in done in the LexRank technique.

V. COMPARATIVE ANALYSIS

It has been noted that in multiple document summarization, an extractive summary may turn out to be more biased depending on the content of the documents involved. [15]

The following Table 1 depicts the differences between extractive and abstractive summarization.

Table 1: Comparison between Extractive and Abstractive Summarization

EXTRACTIVE SUMMARIZATION	ABSTRACTIVE SUMMARIZATION
Since it is purely dependent on the sentences/words extracted from the source text, Incoherency might be found in the abstract formed.	It focuses on the natural language generation of the conceptualized database of the original text. Thus, Abstract formed is coherent.
Suited for both long and short text.	Suited for short text only, Since performance decreases with the size of the source text.[2]
Implementation of techniques used for extractive summarization is not Complex.	Abstractive summarization is more complex than extractive abstraction as it has challenges like Semantic representation, Inference, and Natural Language Generation.

Table 2: Comparison between different types of Extractive Summarization

CENTROID BASED SUMMARIZATION	tf * idf factor of the words of a sentence determines its salience.
GRAPH BASED SUMMARIZATION	Similarity factor of the sentence with respect to central sentences is represented by the degree of the corresponding node in cosine similarity graph and is used to compute the salience of the sentence. It gives better results than centroid based summarization as it addresses the information Subsumption [13].
LEXRANK	Similarity factor of the sentence as well as of all the related sentences are considered to rank the sentences according to their salience. The results shown by this approach are better than graph based summarization because it rules out the high value of idf scores from improving the overall score of a sentence irrelevant to the main theme of text.[13]

VI. CONCLUSION

This paper has reviewed techniques for Automatic Text Summarization. According to our study abstraction gives better results than extraction. However, abstraction proves to be more difficult than extraction and a summary based on abstraction may differ based on opinions. Future work can be done to develop a technique combining the benefits of abstractive and extractive approaches thus achieving a compact and complete summarization. Furthermore, the study reveals that LexRank delivers the best results among the techniques studied in extractive summarization. Firstly, LexRank on comparison with centroid based summarization shows that LexRank is better since it takes information Subsumption into account. Secondly, the study also reveals that LexRank performs better than Graph based Summarization. LexRank prevents the inclusion of sentences which are irrelevant to the main theme of the text whereas Graph based Summarization fails to do so. LexRank prevents the high idf scores pertaining to sentences being irrelevant to the text from improving the overall score of the sentence.

The concept of LexRank is similar to a general purpose graph based ranking algorithm TextRank which is commonly used in Natural Language Processing tasks.

VII. REFERENCES

[1] Moens MF, Angheluta R, Dumortier J. Generic technologies for single-and multi-document summarization. Information Processing & Management. 2005 May 31;41(3):569-86.

[2]Hahn U, Mani I. The challenges of automatic summarization. Computer. 2000 Nov;33(11):29-36.

[3]Luhn HP. The automatic creation of literature abstracts. IBM Journal of research and development. 1958 Apr;2(2):159-65.

[4]Edmundson HP. New methods in automatic extracting. Journal of the ACM (JACM). 1969 Apr 1;16(2):264-85.

[5]Kupiec J, Pedersen J, Chen F. A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval 1995 Jul 1 (pp. 68-73). ACM.

[6] Ferreira R, de Souza Cabral L, Lins RD, e Silva GP, Freitas F, Cavalcanti GD, Lima R, Simske SJ, Favaro L. Assessing sentence scoring techniques for extractive text

summarization. Expert systems with applications. 2013 Oct 15;40(14):5755-64.

[7] Aizawa A. An information-theoretic perspective of tf–idf measures. Information Processing & Management. 2003 Jan 31;39(1):45-65.

[8]Atif K, Naomie S. A review on abstractive summarization methods. Journal of Theoretical and Applied Information Technology. 2014;59(1).

[9]Barzilay R, McKeown KR, Elhadad M. Information fusion in the context of multi-document summarization. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics 1999 Jun 20 (pp. 550-557). Association for Computational Linguistics.

[10]Genest PE, Lapalme G. Fully abstractive approach to guided summarization. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 2012 Jul 8 (pp. 354-358). Association for Computational Linguistics.

[11]Greenbacker CF. Towards a framework for abstractive summarization of multimodal documents. In Proceedings of the ACL 2011 Student Session 2011 Jun 19 (pp. 75-80). Association for Computational Linguistics.

[12]Moawad IF, Aref M. Semantic graph reduction approach for abstractive Text Summarization, International Conference on Computer Engineering and Systems (ICCES), 2012 Seventh International Conference on 2012 Nov 27 (pp. 132-138). IEEE.

[13]Erkan G, Radev DR. LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research. 2004; 22:457-79.

[14]Radev DR, Blair-Goldensohn S, Zhang Z. Experiments in single and multi-document summarization using MEAD. Ann Arbor. 2001; 1001:48109.

[15]Cheung JC. Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection (Doctoral dissertation, UNIVERSITY OF BRITISH COLUMBIA).