Automatic Text Summarization: A review

Naima Zerari, Samia Aitouche, Mohamed Djamel Mouss, Asma Yaha
Automation and Manufacturing Laboratory
Department of Industrial Engineering
Batna 2 University
Batna, Algeria
Email: n.zerari@yahoo.fr, samiaaitouche@yahoo.fr, d_mouss@yahoo.fr, yahaasma@gmail.com

***Abstract*—As we move into the 21st century, with very rapid mobile communication and access to vast stores of information, we seem to be surrounded by more and more information, with less and less time or ability to digest it. The creation of the automatic summarization was really a genius human solution to solve this complicated problem. However, the application of this solution was too complex. In reality, there are many problems that need to be addressed before the promises of automatic text summarization can be fully realized. Basically, it is necessary to understand how humans summarize the text and then build the system based on that. Yet, individuals are so different in their thinking and interpretation that it is hard to create "gold-standard" summary against which output summaries will be evaluated. In this paper, we will discuss the basic concepts of this topic by giving the most relevant definitions, characterizations, types and the two different approaches of automatic text summarization: extraction and abstraction. Special attention is devoted to the extractive approach. It consists of selecting important sentences and paragraphs from the original text and concatenating them into shorter form. Broadly, the importance of sentences is decided based on statistical features of sentences. This approach avoids any efforts on deep text understanding. It is conceptually simple and easy to implement.**

Keywords- Text summarization; Automatic text summarization; Abstractive approach; Extractive approach; Natural language processing.

I. INTRODUCTION

The rapid evolution of WWW has made huge quantity of documents on a variety of topics available to the users [1][2]. To exploit these documents effectively, it is required to be able to get a summary of them. However, it is very difficult for humans to create a hand written summary of the entire available document. Automatic Text Summarization (ATS) provides a solution to this information overload problem [2]. Hence, ATS has become an important and timely tool for user to quickly understand the large volume of information [3]. The automatic summarization included in language processing field, is the process of dealing with a large amount of information by comprising only the essential ones. It often occurs in everyday communication and it is an important and professional skill for some people. Automatic text summarization aims at providing a condensed representation of the content according to the information that the user wants to get [4]. With document summary available, users can easily decide its relevancy to their interests and acquire desired documents with much less mental loads involved. [5].

Furthermore, the goal of automatic text summarization is to condense the documents into a shorter version and preserve important contents [3]. Text Summarization methods can be classified into two major methods extractive and abstractive summarization [6].

The rest of the paper is organized as follows: Section 2 is about text summarization, precisely the definition of the summary; Section 3 describes the automatic text summarization; Section 4 depicts the models of automatic text summarization; Section 5 defines the summaries characteristics; Section 6 presents a brief review of the two text summarization methods and finally Section 7 concludes this paper and outlines the envisaged research work.

II. TEXT SUMMARIZATION

The human being needs a summary mainly because it reduces reading time and it makes the selection process easier during the search of document process.

Text summarization can be used by various applications; for instance researchers need a summary for deciding whether to read the entire document or not and for summarizing information searched by user on Internet. Summarizing documents involves cognitive effort from the summarizer: different fragments of a text must be selected, reformulated and assembled according to their relevance. The coherence of the information included in the summary must also be taken into account [7]. Thus, text summarization, the reduction of a text to its essential content, is a task that requires linguistic competence, world knowledge, and intelligence [7]. The subfield of summarization has been investigated by the Natural Language Processing (NLP) community for nearly the last half century. Radev et al [8] define a summary as: “a text that is produced from one or more texts that convey important information in the original text, and that is no longer than half of the original text(s) and usually significantly less than that”. This simple definition captures three important aspects that characterize research on automatic summarization [8]:

 ISBN: 978-1-61208-542-5

- Summaries may be produced from a single document or multiple documents.
- Summaries should preserve important information.
- Summaries should be short.

The summary done by means of a computer, i.e., automatically, is called Automatic Text Summarization.

III. AUTOMATIC TEXT SUMMARIZATION

Automatic text summarization is the technique which compresses a large text to a shorter text which includes the important information. The computer program is given a text and it returns a summary of the original text. This is done by reducing redundancy of the text and by extracting the essence of the text [9]. Generally, a summary should be much shorter than the source text. This characteristic is defined by the compression rate, which measures the ratio of length of summary to the length of original text [3]. The first effort on automatic text summarization system was made in the late 1950. This automatic summarizer selects significant sentences from the document and concatenates them together [3]. Currently automatic text summarization has benefited from the expertise of a range of fields of research: information retrieval and information extraction, natural language generation, discourse studies, machine learning and technical studies used by professional summarizers [7]. Summaries can be divided in two main categories: extractive and abstractive.

An abstractive summarization tries to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to study the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the salient information from the original text document [6]. This method is the more difficult and it is poorly practical. It is highly complex as it needs extensive natural language processing.

An extractive summarization consists of selecting important sentences or paragraphs from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences [6]. This method is fairly applicable and it usually gives reasonable result. Therefore research community is focusing more on extractive summaries, trying to achieve more coherent and meaning full summaries. Several work have been presented in this context such as: Othman et al. [10] who described the contributions made in text summarization field and presented a comparative study of Text Summarization Techniques. Gupta and Lehal [6] presented a survey of Text Summarization, extractive techniques, specifying that the biggest challenge for text summarization, is to summarize content from a number of textual and semi structured sources, including databases and web pages, in the right way. Saranyamol and Sindhu [11] presented a survey describing different approaches of the automatic text summarization process and made an analysis of different methods. Khan and Salim [3] proposed a survey on abstractive text summarization methods and concluded that most of the abstractive summarization methods produce highly coherent, cohesive, rich information and less redundant summary. Munot and Govilkar [12] discussed in details the two main categories of text summarization methods and also presented a taxonomy of summarization systems, statistical and linguistic approaches for summarization. Sariki et al [2] proposed a system to generate a summary of a single document, specifying the keywords and adjusting the length of the final summary to produce. The proposed system has been improved a lot in accuracy. The authors precise also that the generated summary can be visualized in the form of a Power Point presentation (PPT), thus making it easy for the user to create an effective classroom presentation. So, they propose to extend their work to multiple documents in future. Chandra et al [5] proposed K-mixture semantic relationship significance (KSRS) approach. It is a statistical approach to text summarization. The proposed approach combines the K- mixture term weighting scheme, based on a mathematical (probabilistic) ground, and the linguistic technique. This latter explores term relationships by finding the semantic relationship significance of nouns that signifies term and sentence semantics. The authors specified that the proposed approach, KSRS, performs better and consequently its feasibility in text summarization applications is justifiable. Also, they specified that its use allows the choice of a lower summary proportion without worrying about the performance deterioration.

IV. AUTOMATIC TEXT SUMMARIZATION MODELS

Depending upon the number of documents accepted as input by a summarization process, automatic text summarization can be categorized as single document summarization and multi-document summarization as shown in Fig. 1 below.

In the model Single Document Text Summarization, a summary is produced from single input document. The single document summarization process flow can be depicted in Fig. 2. However, in Multi Document Text Summarization, a summary is produced from multiple input documents dealing with the same topic as illustrate in Fig. 3. In 1995, Radev and McKeown [13] were the first to develop a system for generating summaries of multiple documents. Multidocument summarization is one of the major challenges in current summarization systems because the task of summarizing multiple documents is more difficult than the task of summarizing single documents where the redundancy [1] is the main problems in summarizing multiple documents.

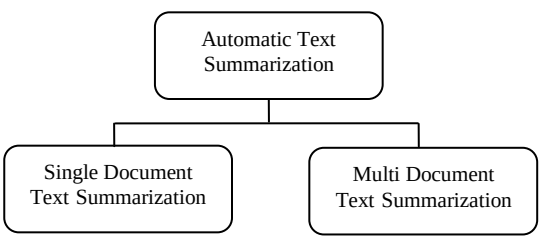

Figure 1. Automatic Text Summarization Models

 ISBN: 978-1-61208-542-5

V. CHARACTERISTICS OF SUMMARIES

The summary is characterized by various features cited below [8]:

1. Language: designates the language of the input; it can be monolingual or multilingual.
2. Genre: represents scientific article, report, news or other.
3. Type of document: specifies the type of the document used as an input; it can be classified into two types:
   a. Single document summarizes: creates a summary from one document.
   b. Multiple documents summary: creates a summary from a number of related documents summarization (more than one document). The distinct characteristic that makes multi document summarization rather different from single document is the use of multiple sources of information that overlap and supplement each other, being contradictory. So the fundamental tasks do not consist just on identifying and coping with redundancy across documents, but also ensuring that the final summary is both coherent and complete.
4. Domain: Corresponds to the domain of summarization such as science, technology, literature, law, etc. It is defined by two types:
   a. Restricted summary: provides summary on restricted domain.
   b. Unrestricted summary: applies for all type of documents. So, there is not dependence on the domain and can be used by any type of user.
5. Type of information: Signifies the type of information used, it encloses two types:
   a. Background information: teaches about the topic.
   b. New information summary**:** provides just the newest facts, assuming the reader is familiar with the topic.
6. Audience: designates the method used to write a summary, defined by two types:
   a. Generic summary: provides the author's point of view. Generic summarization purpose is to summarize all texts regardless of its topic or domain; i.e., generic summaries make no assumptions about the domain of its source information and view all documents as homogenous texts [14].
   b. Query based summary: focuses on material of interest to the user.
7. Function: Signifies the type of the function used to transform the document to a summary, and it covers three types:
   a. Informative summary: reflects the content of the original text.
   b. Indicative summary: merely provides an indication of what the original text was about.
   c. Evaluative summary: evaluates the subject matter of the source, expressing the abstractor's views on the quality of the work of the author.

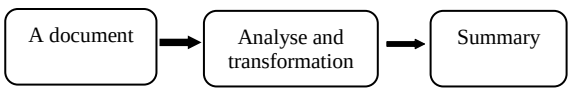

Figure2. Single Document Text Summarization

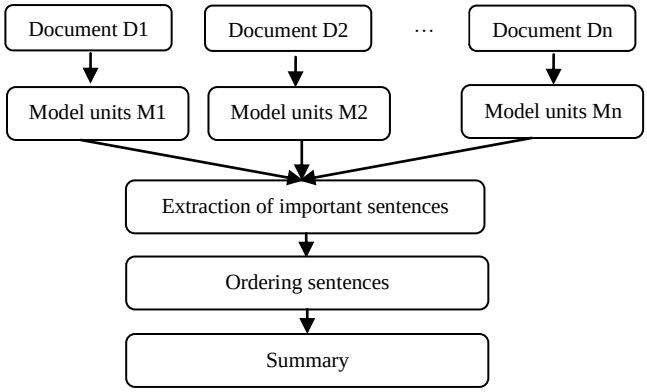

Figure3. Multi Document Text Summarization

VI. METHODS OF SUMMARIZING

The output of summary can be of two types: Extractive summaries and Abstractive summaries. Extractive summaries are produced by extracting the whole sentences from the source text. The importance of sentences is determined based on statistical and linguistic features of sentences [9]. Abstractive summaries are produced by reformulating sentences of the source text. The principle of abstractive summarizer consists to understand the main concepts in a document and then convey those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and terms to best describe it by generating new shorter text that conveys the most significant information from the original text document [9].

A. Extractive Method

Extractive approach purpose is to create the summary by extracting the important sentences from the original document [2]. The extracted sentences will be then grouped to produce a summary with maintaining the order as in the original document and without changing the source text [11]. Most of the work in text summarization has focused on extractive summarization because it is conceptually simple and easy to be implemented. Generally, there are three types of approach to extract sentences in summary generation: the statistical, the linguistic and machine learning approach [10].

1) Linguistic Approach

This technique involves knowledge of the language so that the computer can analyze the sentences semantically and then decide what sentences to choose considering the position of the subject, verb and the noun [10]. It is more difficult than statistical methods.

2) Machine Learning Approach

A Machine Learning (ML) approach is useful where a collection of documents and their corresponding reference extractive summaries are available [15]. The ML aims at learning from a training model in order to determine the appropriate class where an element belongs to. The

sentences of each document will be representing by means of vectors of features extracted from the text [15][14]. Thus, the goal of training model is to classify sentences in two categories: sentence labelled as "summary sentence" when it belong to the reference summary or as "non-summary sentence" other than. This process of learning from the collection of documents and its summaries allow the use of the trained model to produce an extractive summary when a new document is given to the system [14]. Some ML methods used for single document will be described.

A. Text Summarization with Neural Networks

This method involves neural network training to identify the type of sentences that must be inserted in the summary. The neural network learns the patterns that are essential in sentences and that should be included in the summary. Generally, this method uses Feed forward neural network architecture with three layers [11].

B. Text Summarization with Naive Bayes

One of the early works that integrated machine learning was the use of Naive Bayes classifier for learning from the data in 1995 [14]. In this method, the classification function namely naïve- bayes is used to categorize_each sentence as worthy of extraction or not [16][17].

3) Statistical Approach

In Statistical technique, the summary is created without understanding, but rather depends on the statistical distribution of certain properties [10]. This technique aims at deriving weights of key terms and determine the sentence importance by the total weight the sentence contains [5].

- *Statistical Technique Steps*

The statistical technique is realized in the following different steps:

a. Pre-processing
b. Analyzing

a. Pre-processing**:** is the initial step of loading the given text into the proposed system and decomposing it into its constituent sentences (takes a raw text as an input and applies some basic routines to transform or eliminate textual elements that are not useful in further processing of textual data). Normalization is the method of converting the text into normalized form by performing processes, such as case-folding, tokenization, stop word removal and stemming. Thus, the Pre-Processing steps are [2][18]:

- Case-folding ;
- Tokenization ;
- Stop word removal ;
- Stemming.

▪ Case-Folding**:** is the process of converting the given text into lower case text in order to avoid repetition of the same word in different cases. This helps the system to distinguish similar terms and improves its accuracy [2][18].

▪ Tokenization: is the process of splitting text into sentence and each sentence into words. For sentence segmentation, dot is taken as separator and for words space is taken into account [2][18].

▪ Stop word removal**:** is the process of removing the stop words, i.e., words which are of less semantic information. Words which are very common and occur in a large majority of the documents but do not include much semantic information are termed as stop words, such as: "the", "by", "a", "an", etc.

Categorization is only based on feature terms and not on full stops, commas, colons, semicolons, etc. So they are removed from the document and will not be stored in the signature file for further process [2][18].

Stemming: The objective of this process is to obtain the stem or radix of each word (in general, a text document contains repetitions of the same word with variations), which emphasize its semantics [15]. It deals with syntactically-similar words, such as plurals, verbal variations, etc. [15]. The purpose of this procedure is to obtain the stem or radix of each word, which emphasize its semantics [15]. Stemming can be of two types [2]:

- Derivational Stemming.
- Inflectional Stemming.

Derivational stemming creates new words from existing words, e.g., "Finalize-Final", "Useful-Use", "Musical-Music", etc. However, Inflectional stemming confines normalized words to grammatical variants like past tense or present tense or singular or plural form, e.g., "Management-Manage", "Classification-Classify", "Payment-Pay", etc. [2][18].

b**.** Analyzing: This stage has traditionally been decomposed into three steps [2][18]:

- Ranking: Conception of the structure of analyzing using to summarize.
- Selection**:** Transformation by using a function "Statistic function".
- Ordering**:** ordering the new statements for make an understandable summary.

- *Methods of Statistical Technique*

Scoring is the process of assigning a score for each sentence to determine its importance in the summary [2]. Text summarization identifies and extracts key sentences from the source text and concatenates them to form a concise summary. Importance of a sentence can be decided by several methods, such as:

▪ *TF-IDF method (Term Frequency-Inverse Document Frequency)*

This method introduced in 1989 [19]. The term frequency (TF) contributes to the similarity strength as the number of word occurrences is higher. Whereas, the inverse document frequency (IDF) regards low frequency words inversely contributes to higher value to the measurement [19]. The purpose of tf-idf is to reduce the weightage of frequent occurring words by comparing its proportional frequency in the document collection. This property has

made the tf-idf to be one of the commonly used terminologies in extractive summarization [14].

- Cue-Phrase Method

Words that would have positive or negative effect on the respective sentence weight to indicate significance or key idea [3], such as cues: “in summary”, “in conclusion”, ”the paper describes”, ”significantly”.

- Title Method

This method states that sentences that appear in the title are considered to be more important and are more likely to be included in the summary. The score of the sentences is calculated as how many words are commonly used between a sentence and a title. Title method cannot be effective if the document does not include any title information [12].

- Location Method

It relies on the intuition that important sentences are located at certain position in text or in paragraph, such as beginning or end of a paragraph [3]. Therefore, important information in a document is often covered by writers at the beginning of the article. Thus the beginning sentences are assumed to contain the most important content [11].

- Sentence length

Very short sentences are usually not included in summary as they convey less information. Very long sentences are also not suitable to represent a summary [20].

- Proper noun

Sentences containing proper noun representing a unique entity suchlike name of a person, organization or location are considered important to the document [20] [14].

B. Abstractive Method

Abstractive text summarization method is intended to produce important information about the document in a new way, by interpreting and examining the source text and then creating a concise summary, closer to what a human might generate. The summary will contain compressed sentences or may include some novel sentences not present explicitly in the original source text [21][22][23]. It produces an organic summary with a logic structure clearer and more accurate as compared to the summaries produced by extractive approach [12]. However, this method is difficult because it uses linguistic approach to understand the original text [12] and needs deep understanding of the NLP tasks. It is broadly classified in two categories: Structured based approach and Semantic based approach [3].

1) Structured Based Approach

Structured based approach encodes most important information from the document(s) through cognitive schemas [3][11]. Different methods can be used by Structured Based Approach, such as Tree based method, Template based method, ontology based method, lead and body phrase method and Rule based method [3] as illustrated in Fig. 4.

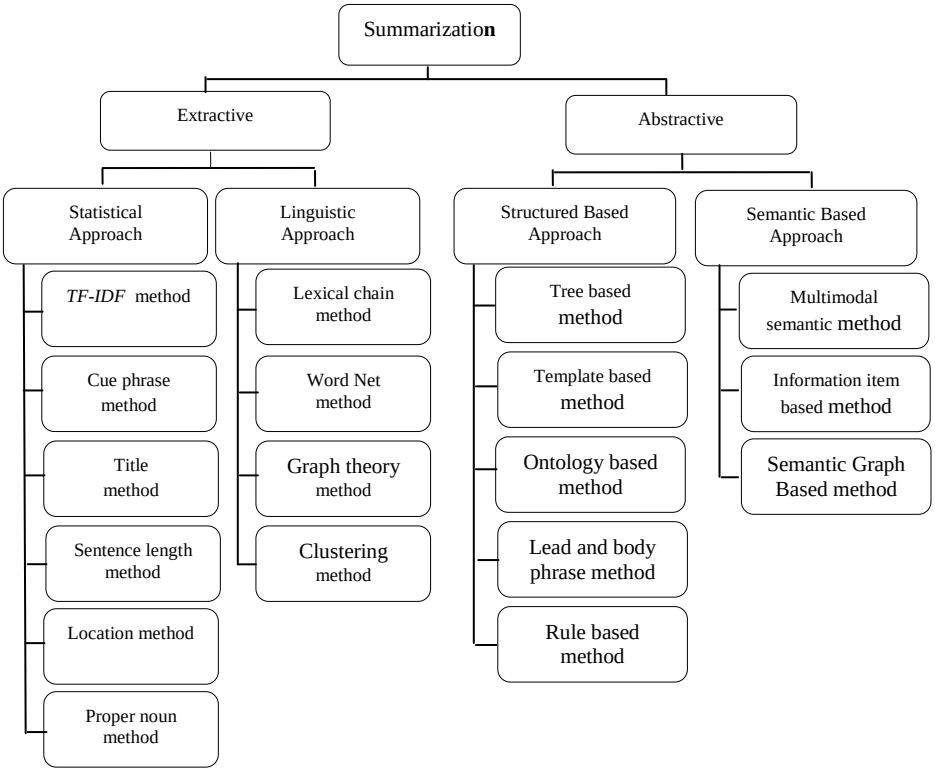

Figure 4. Principles Approaches used in Automatic Text Summarization

2) Semantic Based Approach

In Semantic based approach, a semantic representation of document(s) is used to feed into natural language generation (NLG) system. This method focus on identifying noun phrases and verb phrases by processing linguistic data [3] [11]. Various methods can be used by Structured Based Approach suchlike Multimodal semantic model, Information item based method and Semantic Graph based method [3] as presented in Fig. 4 above.

VII. CONCLUSION AND FUTURE RESEARCH

Nowadays, the need of automatic text summarization has augmented due to the rapid increase in number of information on the Internet. Therefore, it is too difficult for users to manually summarize those large online documents. Automatic text summarization solves this problem. It represents one of the natural language processing applications and is becoming more popular for information condensation. It allows getting the important information while dealing with large collection of documents. A good automatic summary captures the essence of a long work in a brief informative statement that can be read and digested quickly. This solution can be developed using either extractive or abstractive approaches that both aimed at analyzing the texts and generalizing summaries. Text summarization by abstractive approach is stronger because it produces summary which is semantically related but difficult to generate. However, text summarization by extractive approach is easier for the human to program and for the computer to understand. This review mainly focused on the fundamental concepts and approaches related to automatic text summarization and its most important characterization. Therefore, much discussion revolves around the extractive approach due to its great use. However, there are a number of limitations pertaining to this approach that is, its sentences can be extracted out of the context and anaphoric references can be broken. Thus, the main aim of this research work is to understand the text summarization process for developing an automatic text summarization system with great accuracy as future work. This objective can be achieved by applying a hybrid method of statistical approach.

ACKNOWLEDGMENT

This research is supported by Automation and Manufacturing Laboratory of Industrial Engineering Department of Batna-2 University.

REFERENCES

[1] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques : a survey," *Artif. Intell. Rev. Springer Sci. Media Dordr.*, vol. 47, no. 1, pp. 1–66, 2016.

[2] T. P. Sariki, B. Kumar, and R. Ragala, "Effective classroom presentation generation using text summarization," *Comput. Technol. Appl.*, vol. 5, no. August, pp. 1–5, 2014.

[3] A. Khan and Naomie Salim, "A Review On Abstractive Summarization Methos," *J. Theor. Appl. Inf. Technol.*, vol. 59, no. 1, 2014.

[4] F. Kiyoumarsi, "Evaluation of Automatic Text Summarizations based on Human Summaries," *Procedia - Soc. Behav. Sci.*, vol. 192, pp. 83–91, 2015.

[5] M. Chandra, V. Gupta, and S. K. Paul, "A Statistical Approach for Automatic Text Summarization by Extraction," in *International Conference on Communication Systems and Network Technologies*, 2011, pp. 268–271.

[6] V. Gupta, Gurpreet Singh Lehal, and G. S. Lehal, "A Survey of Text Summarization Extractive techniques," *J. Emerg. Technol. Web Intell.*, vol. 2, no. 3, pp. 258–268, 2010.

[7] J.-M. Torres-Moreno, *Automatic Text Summarization*. British Library Cataloguing-in-Publication Data. ISTE Ltd, John Wiley & Sons, Inc., 2014.

[8] D. Das and A. F. Martin, "A Survey on Automatic Text Summarization," *Lit. Surv. Lang. Stat. II course C. 4*, pp. 192–195, 2007.

[9] P. Shah and N. P. Desai, "A Survey of Automatic Text Summarization Techniques for Indian and Foreign Languages," in *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016.

[10] B. M. M. Othman, M. Haggag, and M. Belal, "A Taxonomy for Text Summarization," *Inf. Sci. Technol.*, vol. 3, no. 1, pp. 43–50, 2014.

[11] C. S. Saranyamol and L. Sindhu, "A Survey on Automatic Text Summarization," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 6, pp. 7889–7893, 2014.

[12] N. Munot and S. S. Govilkar, "Comparative Study of Text Summarization Methods," *Int. J. Comput. Appl.*, vol. 102, no. 12, pp. 33–37, 2014.

[13] L. Suanmali and N. Salim, "Literature Reviews for Multi-Document Summarization.", 2008.

[14] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A Review on Automatic Text Summarization Approaches," *J. Comput. Sci.*, 2016.

[15] J. Neto, A. Freitas, and C. Kaestner, "Automatic Text Summarization Using a Machine Learning Approach," *Adv. Artif. Intell. Bittencourt, G. G.L. Ramalho, Springer-Verlag Berlin Heidelb.*, pp. 205–215, 2002.

[16] S. Suneetha, "Automatic Text Summarization : The Current State of the art," *Int. J. Sci. Adv. Technol.*, vol. 1, no. 9, pp. 283–293, 2011.

[17] N. Bhatia and A. Jaiswal, "Literature Review on Automatic Text Summarization : Single and Multiple Summarizations," *Int. J. Comput. Appl.*, vol. 117, no. 6, pp. 25–29, 2015.

[18] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," *Tech. – Int. J. Comput. Sci. Commun. Technol.*, vol. 2, no. 1, pp. 325–335, 2009.

[19] M. Haque, S. Pervin, and Z. Begum, "Literature Review of Automatic Multiple Documents Text Summarization," *Int. J. Innov. Appl. Stud.*, vol. 3, no. 1, pp. 121–129, 2013.

[20] Y. J. Kumar and N. Salim, "Automatic multi document summarization approaches," *J. Comput. Sci.*, vol. 8, no. 1, pp. 133–140, 2012.

[21] M. Bhide, "Single or Multi-document Summarization Techniques," vol. 4, no. 3, pp. 375–379, 2016.

[22] S. Haiduc, J. Aponte, L. Moreno, and A. Marcus, "On the use of automated text summarization techniques for summarizing source code," *Proc. - Work. Conf. Reverse Eng. WCRE*, pp. 35–44, 2010.

[23] M. S. Patil, M. S. Bewoor, and S. H. Patil, "A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 1584–1586, 2014.

 ISBN: 978-1-61208-542-5