 ISBN: 978-81-927230-0-6

# Survey on Extractive Text Summarization Approaches

M S Patil[#1], M S Bewoor [*2], S H Patil [#3]

[#1]*M.Tech Computer Department, Bharati Vidypeeth University College of Engineering Pune,India*

[#2]*Associate Professor Computer Department, Bharati Vidypeeth University College of Engineering Pune,India*

[#3] *Professor Computer Department, Bharati Vidypeeth University College of Engineering Pune,India*

[1]madhsp.patil@gmail.com

[2]msbewoor@bvucoep.edu

[3]shpatil@bvucoep.edu

***Abstract*— Due to increasing use of internet and online technologies or online data, there is vast increase in the electronic documents. When a data is being retrieved from such a huge collection of electronic documents, hundreds and thousands of documents are retrieved. Hence, for user, it is not possible to read all the retrieved documents. Also, these documents contain redundant information. In such situation summarization proves to be very useful that summarizes the retrieved documents. This has lead to intensive research in the area of automatic text summarization. It is widely used in other fields like natural language processing and machine learning. Text summarization is the process of generating summary of one or more documents which conveys information present in the documents and is usually less than the original documents. This research paper provides an overview of the extractive techniques for text summarization. Two main approaches focused are clustering techniques and machine learning technique (SVM).**

***Keywords*— abstractive summarization, clustering, extractive summarization, pre-processing, SVM**

## I. INTRODUCTION

Due to increasing electronic document, text summarization is gaining much importance nowadays. This leads to the need of automatic text summarization. Users trying to retrieve the documents or information face the problem of responses of hundreds or thousands of retrieved documents, as the number of documents on internet has increased. It is impossible to summarize such a large number of documents retrieved, without using automatic summarization. In addition to this, one more problem faced is redundancy in the information presented by the retrieved documents. A solution for all these problems is to summarize these documents automatically. Text summarization summarizes the text documents thereby solving the problems.

Text summarization is the process of generating summary of one or more documents which conveys information present in the documents and is usually less than the original documents. It produces summary that reduces the redundancy in the text thereby preserving the information. In short summarization may be seemed to cover three important aspects:

- Summaries thus generated must be short.
- Summaries must have less redundant information.
-They must preserve important information.

Summarization is an important activity in the analysis of a huge volume text documents. Goal of summarization is to convey the main concept of the document in such a way that the user will not have to waste time in reading the document and also it is well understood. Summary generated by the summarization technique is nothing but a text extracted from original document that explains the information present.

Summarization is broadly classified into two types:

**Extractive:** Extractive summarization methods simplify the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. This type of summarization picks out the most relevant sentences in the document thereby maintaining a low redundancy in the summary.

**Abstractive:** Abstractive summarization may compose novel sentences, unseen in the original sources.

.

## II. RELATED WORK

This section gives the overview of the research work carried out related to the summarization. This overview mainly focuses on the clustering techniques and the machine learning approaches.

Kamal Sarkar in [3] has proposed an approach to Sentence Clustering-based Summarization. In this author has considered three important factors: sentence clustering, ordering cluster and selecting representative sentences from clusters. Here author has measured the importance of cluster on the important words present in the cluster. Top n clusters are selected among the clusters ordered in the decreasing order. For summary each representative sentence is selected from each cluster. This selection proceeds till summary of fixed size is obtained.

Jing and McKeown in [1] has discussed about the summarization approach based on statistical method. In this first sentences are extracted that are most important. These sentences are then concatenated and modified as needed. In this pre-processing step is discussed that contains sentence reduction, sentence combination, syntactic transformation and lexical paraphrasing are the sub phases. Sentence reduction removes extraneous phrases. Sentence combination combines sentences depending on their context. Syntactic transformation is the altering of the grammatical structure of the sentences. Lexical paraphrasing is replacing phrases with their paraphrases. Reordering is reordering of the selected sentences in order to make the comprehensible summary.

In [7] author has proposed the analysis of Parts of Speech Tagging for a feature term based text summarization technique. Based on identification and extraction of important sentences in the document a new approach of generating summary for a given input document is discussed. For this, selective terms are obtained from the extracted terms and qualitative summary is built with appreciable compression ratio. Many users have limited knowledge regarding the appropriateness and relevance of the information in the web pages because of absence of contextual and discourse awareness in today's web. In this paper, author has provided text summarization approach as a solution to this problem. This approach reduces the time required to find the web document having relevant and useful data.

Michael Steinbach, George Karypis, Vipin Kumar in [5] presented the results of the experiments studied consisting of some common document clustering techniques. In this, comparison of the two main approaches to document clustering is done, agglomerative hierarchical clustering and K-means. Hierarchical clustering is often considered as the better quality clustering approach. Its quadratic time complexity is its limitation. In contrast, K-means have a linear time complexity for number of documents. But they produce inferior clusters. To get best of both sometimes agglomerative hierarchical and K-means approaches are combined. Many times though the nearest neighbors of a document are of different classes agglomerative hierarchal clustering often place the documents of the same class in the same cluster, even at the early stages of clustering techniques. The mistakes occurred cannot be fixed because of the way that hierarchical clustering works. In this author has discussed that to find clusters that corresponds the desired document classes K-means fails sometimes. Also author has mentioned in this that K-means general approach suits better to documents than that of the agglomerative clustering.

Vibekananda Dutta, Krishna Kumar Sharma, Deepti Gahalot [9] has presented an overview of method for soft approaches of an optimal fuzzy document clustering algorithm as compare to the hard approaches. Experiments conducted by the author consist of application of K-means and Fuzzy c-means on document sets. In the comparison made by the author Fuzzy Clustering approach is potentially suitable for document clustering. In this paper author has conclude that the Fuzzy clustering gives better results than K-Means clustering algorithm.

Roma J, M S Bewoor, S H Patil in [7] has presented the comparision of various techniques of text summarization. Evaluation of different summaries has been done basedof some qualitative and quantitative techniques. Results show the comparison of summary generated by the various techniques.

Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda in [10] has proposed extraction of sentences that contain important information from a document. According to author it is the technique for text summarization is the key to the automatic generation of summaries that will generate similar summaries written by humans. Integration of Heterogeneous pieces of information must be done to achieve such extraction. One approach is parameter tuning by machine learning. It has attracted a lot of attention. This paper proposes a method of sentence extraction based on Support Vector Machines (SVMs). To confirm the method's performance, author has conducted experiments that compare their method to three existing methods. Results on the Text Summarization Challenge (TSC) corpus show that this method offers the highest accuracy. Moreover, it is clarified that the different features effective for extracting different document genres.

In [4] author has compared the performance of neural networks and Support Vector Machines for text summarization. These both techniques have ability to discover nonlinear data. Also they are effective for large datasets. Results of the experiments conducted by author shows that neural network are slower than SVM in large datasets.

## III. Proposed System

Proposed system for Text summarization consists of main three steps pre-processing step, processing step and summary generation. Pre-processing step obtains a structured representation of the original text. Processing step deals with the algorithm that transforms the text into summary. Processing step of proposed system consists of clustering technique cascaded with Support Vector Machine. This will improve the quality of summary generated only by the clustering technique.

## IV. Conclusions

This survey paper concentrates on summarization methods based on extractive approach. Summary generated by this approach consists of selection of most important sentences from original documents. Extractive summarization is used most of the times because it simply extracts sentences and this strategy of extraction has produced satisfactory results in case of large scale applications and also in multi document summarization. Here extractive methods focused are clustering method, machine learning and neural method. This survey will be useful for researchers working on summarization to choose appropriate extracting strategy. This

work can be extended by including the different evaluation methods employed for evaluating summary generated by the different summarization techniques.

## REFERENCES


[1] H. Jing and K. McKeown. "Cut and paste based text summarization". In Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pages 178--185, 2000

[2] Jiaming zhan and Han-tong Loh "Using Redundancy Reduction in Summarization to Improve Text Classification by svms" Journal of information science and engineering 25, 591-601 (2009)

[3] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", *TECHNIA – International Journal of Computing Science and Communication Technologies*, vol. 2, no. 1, Jul. 2009

[4] Keivan Kianmehr, Shang Gao, Jawad Attari, M. Mushfiqur Rahman,Kofi Akomeah, Reda Alhajj, Jon Rokne and Ken Barker "Text Summarization Techniques: SVM versus Neural Networks" Proceedings of iiWAS2009

[5] Michael Steinbach George Karypis Vipin Kumar "A Comparison of Document Clustering Techniques" In KDD workshop on Text Mining, 2002

[6] Neepa Shah, Sunita Mahajan "Document Clustering: A Detailed Review" International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5,October 2012

[7] Roma J, M S Bewoor, S H Patil, "Automation tool for Evaluation of the Quantity of NLP Based Text Summarization and Clustering Techniques By Quantitative and Qualitative Metrics" International Journal of Scientific & Engineering Research 2013

[8] Suneetha Manne Shaik, Mohammed Zaheer Pervez, Dr. S. Sameen Fatima "A Novel Automatic Text Summarization System with Feature Terms Identification" India Conference(INDICON), Annual IEEE 2011

[9] Vibekananda Dutta, Krishna Kumar Sharma Deepti Gahalot Performance Comparison of Hard and Soft Approaches for Document Clustering" International Journal Of Computer Applications March 2012

[10] Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda "Extracting Important Sentences with Support Vector Machines" ACM 2002