**A Review Paper On Text Summarization**

**Arifur Rahman and Ashraful Islam**

**Department of IT ,University of Information Technology and Sciences**

**Vatara , Gulshan-2,Dhaka,Bangladesh**

**Abstract :** Text summarization refers to the technique of extracting or collecting important information from original text and turned that information to summary. Text Summarization is now a days a popular and necessity of many application like search engine, business analysis, market review etc. It is a process of natural language processing. Summarization helps quickly and accurately understand big text without much time consuming. There are two major approaches. One is extraction based summarization and the another one is abstraction based summarization. These two approaches will be discussed detail in this paper.This paper provides an abstract view of the present scenario of our research about text summarization.

**Index Terms –** Text Summarization, Natural Language Processing ,Extractive Summary, Abstractive Summary.

**1. INTRODUCTION**

Text summarization is a process of extracting of collecting important information from original text and present that information in the form of summary. In recent years, the data is become huge and huge. This huge amount of data consume much time to read. Text summarization reduce that time consuming and make a summary of the huge data. The IDC projects that the total amount of data which is digital that circulating annually around the world would sprout from 4.4 zettabytes in 2013 to hit 180 zettabytes in 2025. That’s a huge amount of data. So there must be a machine learning algorithms that can automatically shorten the longer text and make accurate summaries that can be easily understandable. Text summarization reduces reading time , faster the process of researching information and increases the amount of information that can fit in an area. The aim of automatic text summarization is to convert large document into shorter one and store important content. The automatic summarization of text is a well- known task in the field of natural language processing (NLP). significant achievements in text summarization have been obtained using sentence extraction and statistical analysis. Text summarization approaches divided into two groups : extractive summarization and abstractive summarization. An extractive summarization technique consists of selecting vital sentences , paragraphs, etc from the original manuscript and concatenating them into a shorter form. he significance of sentences is strongly based on statistical and linguistic features of sentences. The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method.

**2.FEATURES OF TEXT SUMMARIZATION**

Text summarizers identify and extract key sentences from the source text and concatenate them to form a concise summary.There are some features of text summarization :

A. Term frequency : Salient terms provided by statistics are based on term frequency, thus salient sentences are those words that occur repeatedly. The frequently occurring word increases score of sentences. The most common measure widely used to calculate the word frequency is TF IDF.

B. Location : It depends on the intuition that important sentences are located at certain position in text or in paragraph, such start or end of a paragraph . First and last sentence of paragraph has greater chance to be included in summary.

C. Cue Method : Effect of positive or negativity of word on the sentence weight to indicate importance or key idea such as cues: “in summary”, “in conclusion”, “the paper describes”.

D. Title/Headline word : Words in the title and heading of a document that occur in sentences are positively related to summarization. Words that appear in the title are also indicative of the topic or subject of the document.

E. Sentence length : Keeps in view the size of summary. Generally, very long and very short sentences are also not suitable for summary.

F. Similarity : Similarity can be calculated with linguistic knowledge. It indicates similarity between the sentence and title of the document, and similarity between the sentence and remaining sentence of the document.

G. Proper noun : For document summarization sentences having proper nouns are important. Like ,name of a person, place or organization.

H. Proximity : The distance between text units where entities occur is a determining factor for establishing relations between entities.

**3.TECHNIQUES OF TEXT SUMMARIZATION**

There are two techniques of text summarization. They are extractive summarization and abstractive summarization. These two technique is described in the following section :

A. Extractive Summarization Approach

Extractive text summarization is classified in two ways. They are unsupervised learning and supervised learning.All recent works on unsupervised learning technique for text summarization.

Extractive Summarization

Unsupervised Learning

Supervised Learning

Supervised

Learning

Fig 1 . Extractive Summarization

Unsupervised Learning :

Unvervised learning is a learning that learns from test data that has not been labeled, classified or categorized. The unsupervised approaches do not need user input in deciding the important features of the document , this requires the most sophisticated algorithm to provide compensation for the lack of human. This method is a successful method in text summarization. This method is four type . They are graph based approach, concept based , fuzzy logic based , latent semantic analysis (LSA). Raph based models are extensively used in document summarization since graphs can efficiently represent the document structure. Extractive text summarization using external knowledge from Wikipedia incorporating bipartite graph framework has been used. They have proposed an iterative ranking algorithm (HITS algorithm) which is efficient in selecting important sentences and also ensures coherency in the final summary. The uniqueness of this paper is that it combines both graph based and concept based approach towards summarization task. The fuzzy logic approach mainly contains four components. Defuzzifier , fuzzifier , fuzzy knowledge base and inference engine. The textual characteristics input of Fuzzy logic approach are sentenced length, sentence similarity etc which is later given to the fuzzy system. In concept-based approach, the concepts are extracted from a piece of text from external knowledge base such HowNet and Wikipedia . In the methodology proposed , the importance of sentence is calculated based on the concepts retrieved from HowNet instead of words. A conceptual vector model is built to obtain a rough summarization and similarity measures are calculated between the sentences to reduce redundancy in the final summary. Latent Semantic Analysis(LSA) is a method which excerpt hidden semantic structures of sentences and words that are popularly used in text summarization task. It is an unsupervised learning approach that does not demand any sort of external or training knowledge. LSA captures the text of the input document and excerpt information such as words that frequently occur together and words that are commonly seen in different sentences.

Supervised Learning :

Supervised extractive summarization related techniques are based on a classification approach at sentence level where the system learns by examples to classify between summary and non-summary sentences. There are some type of supervised learning. a Machine Learning algorithm that allows us to map numeric inputs to numeric outputs, fitting a line into the data points called linear regression. A classification algorithm that is widely used when the dependent variable is binary (0 or 1) is binary regression. A Machine Learning framework that gets its effectiveness from introducing non-linearity to linear ML models is in neural networks. a Machine Learning algorithm that uses Margin Maximization in determining the optimal separator line between classes, utilizing the Kernel Trick that support vector machines.

B. Abstractive learning

The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method. The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text just like humans do. Therefore, abstraction performs better than extraction. However, the text summarization algorithms required to do abstraction are more difficult to develop; that’s why the use of extraction is still popular.Abstractive techniques are classified into two categories . They are structured based approach and semantic based approach.

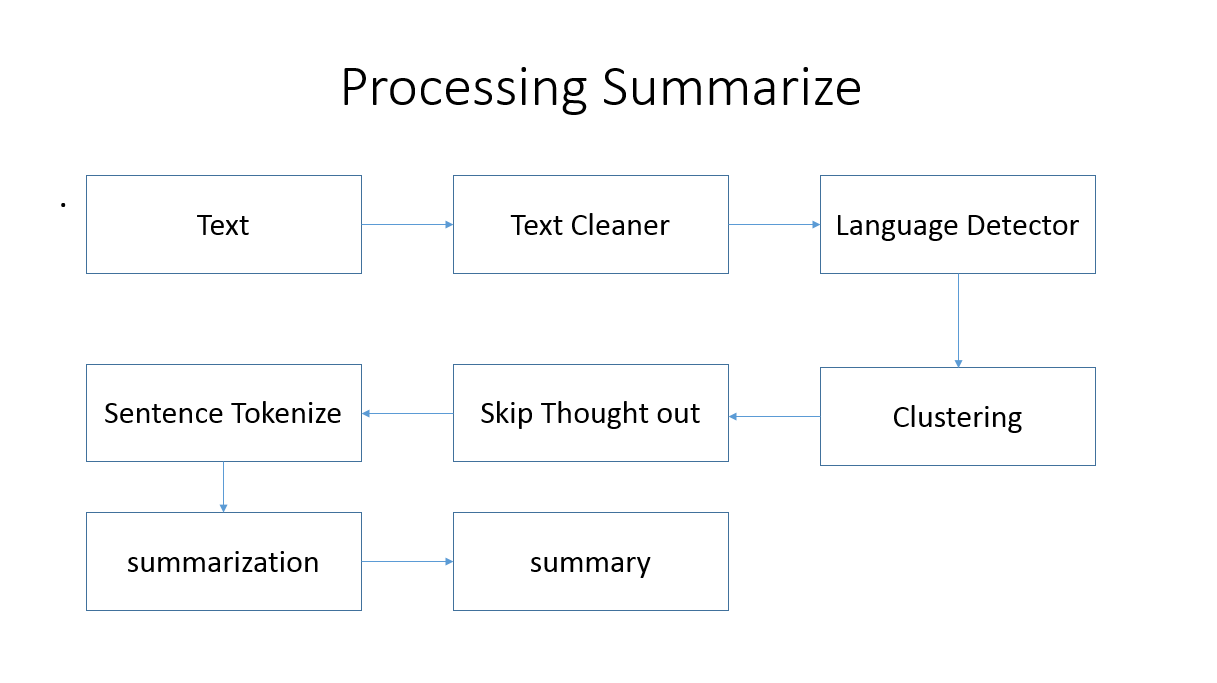
Structured based approach :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Methods** | **Description** | **Advantages** | **Limitation** | **Author & Year** |
| Tree Based Method | -It uses a dependency tree to represent the text of a document. -It uses either a language generator or an algorithm for generation of summary. | - It walks on units of the given document read and easy to summary. | - It lacks a complete model which would include an abstract representation for content selection. | Barzilay and McKeown (1999, 2005) [2], Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, Masaaki Nagata (2014) [17], Tsutomu Hirao, M. Nishino, Y. Yoshida, Jun Suzuki, N. Yasuda, and Masaaki Nagata, (2015) [18]. |
| Template Based Method | -It uses a template to represent a whole document. -Linguistic patterns or extraction rules are matched to identify text snippets that will be mapped into template slots. | -It generates summary is highly coherent because it relies on relevant information identified by IE system. | -Requires designing of templates and generalization of template is to difficult. | Harabagiu and Lacatusu (2002) [2], Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, Raymond Ng (2014) [19]. |
| Ontology Based Method | -Use ontology (knowledge base) to improve the process of summarization. -It exploits fuzzy ontology to handle uncertain data that simple domain ontology cannot. | -Drawing relation or context is easy due to ontology - Handles uncertainty at reasonable amount | -This approach is limited to Chinese news only. - Creating Rule based system for handling uncertainty is a complex task. | Lee and Jian (2005) [2], Meghana viswanath(2006) [9], Ramezani Majid, Feizi Derakhshi Mohammad Reza(2015) [20], R. Ragunath and N. Sivaranjani (2015) [23]. |
| Lead and Body Phrase Method | - This method is based on the operations of phrases (insertion and substitution) that have same syntactic head chunk in the lead and body sentences in | -It is good for semantically appropriate revisions for revising a lead sentence. | -Parsing errors degrade sentential completeness such as grammaticality and repetition. -It focuses on rewriting techniques, and |  |

Semantic based approach :

|  |  |  |  |
| --- | --- | --- | --- |
| **Methods** | **Description** | **Advantages** | **Author & Year** |
| Multimodal semantic model | -A semantic model, which captures concepts and relationship among concepts, is built to represent the contents of multimodal documents. | -An important advantage of this framework is that it produces abstract summary, whose coverage is excellent because it includes salient textual and graphical content from the entire document. | Greenbacker (2011) [2]. |
| Information Item Based Method | -The contents of summary are generated from abstract representation of source documents, rather than from sentences of source documents. -The abstract Representation is Information Item, which is the smallest element of coherent information in a text. | -The major strength of this approach is that it produces short, coherent, information rich and less redundant summary. | Genest and Lapalme (2011) [2], Daniel Mallett, James Elding, Mario A. Nascimento (2004) [15]. |
| Semantic Graph Based Method | -This method is used to summarize a document by creating a semantic graph called Rich Semantic Graph (RSG) for the original document, reducing the generated semantic graph. | - It produces concise, coherent and less redundant and grammatically correct sentences. | Moawad & Aref (2012) [2], Kavita Ganesan, ChengXiang Zhai & Jiawei Han, (2010) [30], Laura Plaza, Alberto Díaz & Pablo Gervás, (2011) [31], Manjula Subramaniam, Prof. Vipul Dalal(2015)[37] |

**4.PROCESS OF TEXT SUMMARIZATION**

Text summarization in natural language processing is treated as a supervised machine learning problem. Introduce a method to extract the merited keyphrases from the source document. For example, you can use part-of-speech tagging, words sequences, or other linguistic patterns to identify the keyphrases. Gather text documents with positively-labeled keyphrase . The keyphrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labeled keyphrases. Train a binary machine learning classifier to make the text summarization. In the test phrase, create all the keyphrase words and sentences and carry out classification for them. We first need to convert the whole paragraph into sentences. The most common way of converting paragraphs to sentences is to split the paragraph whenever a period is encountered. After converting paragraph to sentences, we need to remove all the special characters, stop words and numbers from all the sentences.Then perform tokenization on the text. Tokenization is the process of dividing text into a set of meaningful pieces. Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens become the input for another process like parsing and text mining. Then it checks frequency of the data. After checking the frequency it prepeare data and process the data. .

**5.CONCLUSION**

Text summarization is an interesting machine learning field that is increasingly gaining traction. As research in this area continues, we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. It is growing as sub branch of NLP as the demand for compressive, meaningful, abstract of topic due to large amount of information available on net. text summarization is need and used by business analyst, marketing executive, development, researchers, government organizations, students and teachers. Text summarization has its importance in both commercial as well as research community. As abstractive summarization requires more learning and reasoning, it is bit complex then extractive approach but, abstractive summarization provides more meaningful and appropriate summary compare to extractive. The proposed work does not involve a knowledge base and therefore can be used to summarize articles from fields as diverse as politics, sports, current affairs, and finance. A possible application of this work can be made to make data available on the move on a mobile network by even shortening the sentences produced by our algorithm and then shortening it. Various NLP based algorithms can be used to achieve this objective. This will ensure that the summary produced is to the highest condensed form which can be made in the mobile industry

**REFERENCES**

1.Saranyamol C S, Sindhu L, “A Survey on Automatic Text Summarization”, International Journal of Computer Science and Information Technologies, 2014,Vol. 5 Issue 6.

2. Reeve Lawrence H., Han Hyoil, Nagori Saya V., Yang Jonathan C.,Schwimmer Tamara A., Brooks Ari D.,“Concept Frequency Distribution in Biomedical Text Summarization”, ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, USA,2006.  
3.Blog.mashape.com/list-of-30–summarizer-libraries-and-software.  
4. Khan Atif, Salim Naomie, “A review on abstractive summarization Methods”, Journal of Theoretical and Applied Information Technology, 2014, Vol. 59 No. 1.  
5. Suneetha Manne, Zaheer Parvez Shaik Mohd. , Dr. S. Sameen Fatima, “Extraction Based Automatic Text Summarization System with HMM Tagger”, Proceedings of the International Conference on Information Systems Design and Intelligent Applications, 2012, Vol. 132, P.P 421-428.  
6. Gupta Vishal, “A Survey of Text Summarizers for Indian Languages and Comparison of their Performance”, Journal of Emerging Technologies In Web Intelligence, 2013, Vol. 5, No. 4.  
7. Vishal Gupta,“A Survey of Recent Keywords and Topic Extraction Systems for Indian Languages”, International Journal of Engineering Trends and Technology (IJETT) , 2013, Vol. 6 No. 6  
8. Gupta V. And Lehal G. S., “A Survey of Text Summarization Extractive Techniques”, International Journal of Emerging Technologies in Web Intelligence, 2010, Vol. 2. , pp. 258-268.  
9. Viswanath Meghana, “Thesis: Ontology-Based Automatic Text  
Summarization”, M. Sc Thesis, Vishweshwaraiah Institute of Technology, India, 2009.  
10. Sarkar Kamal, Nasipuri Mita, Ghose Suranjan, “Using Machine Learning for Medical Document Summarization”, International Journal of Database Theory and Application, 2011.  
11.Ladda Suanmali, Salim Naomie, and Mohammed Salem Binwahlan, “Fuzzy Logic Based Method for Improving Text summarization” IJCSIS, 2009.  
12. Ibrahim Imam, Nihal Nounou, Alaa Hamouda, Hebat Allah Abdul Khalek, “Query Based Arabic Text Summarization”, IJCST, 2013, Vol. 4, Issue Spl - 2.  
13. Khosrow Kaikhan, “Text Summarization Using Neural Networks”, Proceedings. 2004 Second IEEE International Conference on Intelligent Systems, 2004, Vol. 1. Mehdi Bazghandi, Ghamarnaz Tadayon Tabrizi and Majid Vafaei Jahan, “Extractive Summarization of Farsi Documents Based On PSO Clustering”, International Journal of Computer Science Issues, 2012, Vol. 9, Issue 4, No 3.  
15. Mallett Daniel, Elding James, Nascimento Mario A., “InformationContent Based Sentence Extraction for Text Summarization”,Proceedings. ITCC 2004. International Conference on, 2004,Vol. 2.  
16. Deshpande Anjali R., Lobo L. M. R. J., “Text Summarization using Clustering Technique”, International Journal of Engineering Trends and Technology (IJETT) , 2013, Vol. 4 Issue8.  
17. Karmakar Surajit, Lad Tanvi, Chothani Hiten, “A Review Paper on Extractive Techniques of Text Summarization”, International Research Journal of Computer Science (IRJCS), Issue 1, 2015, Vol. 2.  
18. Kikuchi Yuta, Hirao Tsutomu, Takamura Hiroya, Okumura Manabu, Nagata Masaaki, ” Single Document Summarization based on Nested Tree Structure” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 315–320, Baltimore, Maryland, USA, 2014.  
19. Hirao Tsutomu, Nishino Masaaki, Yoshida Yasuhisa, Suzuki Jun, Yasuda Norihito, and Nagata Masaaki, “Summarizing a Document by Trimming the Discourse Tree”, IEEE/ACM Transactions On Audio, Speech, And Language Processing, 2015, Vol. 23, No. 11.  
20. Ramezani Majid, Feizi-Derakhshi Mohammad-Reza, “OntologyBased Automatic Text Summarization Using FarsNet”, ACSIJ Advances in Computer Science: an International Journal, 2015, Vol. 4, Issue 2, No.14.  
21. Oya Tatsuro, Mehdad Yashar, Carenini Giuseppe, and Ng Raymond, “A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships”, Proceedings of the 8th International Natural Language Generation Conference, pages 45–53, Philadelphia, Pennsylvania, 2014.  
22. Patil Pallavi D., Kulkarni N.J., “Text Summarization Using Fuzzy Logic”, International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2014, Vol. 1 Issue 3.  
23. Froud Hanane, Lachkar Abdelmonaime and Ouatik Said Alaoui, “Arabic Text Summarization Based On Latent Semantic Analysis To Enhance Arabic Documents Clustering”, International Journal of Data Mining & Knowledge Management Process (IJDKP), 2013, Vol.3, No.1.  
24. Ragunath R. And Sivaranjani N., “Ontology Based Text Document Summarization System Using Concept Terms”, ARPN Journal Of Engineering And Applied Sciences, 2015, Vol. 10, No. 6.  
25. Patil Pallavi D, Mane P M, “A Comprehensive Review on Fuzzy Logic & Latent Semantic Analysis Techniques for Improving the Performance of Text Summarization”, International Journal of Advance Research in Computer Science and Management Studies, 2014, Vol. 2, Issue 11, pg. 476-485.  
26. Dixit Rucha S., Apte S. S., “Improvement Of Text Summarization Using Fuzzy Logic Based Method”, IOSR Journal Of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727, 2012, Vol. 5, Issue 6, PP 05-10.  
27. Fachrurrozi M., Yusliani Novi, and Yoanita Rizky Utami, “Frequent Term based Text Summarization for Bahasa Indonesia”, International Conference on Innovations in Engineering and Technology Bangkok (Thailand), 2013.  
28. Babar S.A. and Thorat S.A., “Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis”, International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2014, Vol. 1 Issue 4.  
29. Megala S. Santhana, Kavitha A., Marimuthu A., “Enriching Text Summarization using Fuzzy logic”, International Journal of Computer Science and Information Technologies, 2014, Vol. 5 Issue 1.30. Ganesan Kavita, Zhai ChengXiang and Han Jiawei, “Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions”, Proceedings of the 23rd International Conference on computational Linguistics (Coling 2010), 2010, pages 340–348.