# Automatic Text Summarization and it's Methods - a Survey

Udit Chauhan[1] and Tarun Tiwari[2]
[1]SRM Institute of Science and Technology , Kattankulathur , Chennai Email: chauhan.udit444@gmail.com
[2]SRM Institute of Science and Technology , Kattankulathur , Chennai Email: tarun.twr0075@gmail.com

*Abstract*— **Text summarization is a very important problem in natural language processing. A conversation between two people or a group of people includes exchange of important information sandwiched between lots of unimportant discussion leading up to it. It therefore becomes necessary to note important key points and essence of the conversation for later reference. Through this paper we aim to highlight the automation of the process. The relevant information is extracted from the conversation using NLP techniques like Decompounding, Entity extraction like regex extraction (SSN, Phn. No), statistical extraction (names, geographical places), Phrase extraction etc. Here in our study we will be mainly focusing on developing a summary for the user through text summarization techniques such as Abstractive text summarization and Extractive text summarization.**

**Keywords: Automatic Text Summarization, Extractive Text Summarization, Abstractive Text Summarization, Natural Language Processing (NLP), Machine Learning**

## I. INTRODUCTION

In the recent years the data has been increased exponentially in the form of text, image, audio, video, graphics, animations etc. Out of all the above textual data has been of immense importance generated from numerous sources. Once such source of text generation is from voice/audio. Today no business can flourish without keeping track on information and important points, agendas discussed in their regular meetings. Hence it is very important to store this useful information safely for future reference so that whenever user wants to know something he can get the same info quickly. But the problem is there is a bulk of data available. Hence to save time and effort Summarizing this stored information and present this useful information in a clear and concise format becomes extremely important. For achieving above task first the voice/audio is converted into textual format using Google speech-to-text Api followed by automatic text summarization. And then replying back to the user queries in voice/text format. The principle techniques used for this process are Natural Language Processing and automatic text summarization.

1. Natural Language Processing: Several of the most commonly used NLP techniques can be implemented in this project. This mainly includes techniques for Sentence Decompunding, Entity extraction using several regular expression for regex extraction like (SSN, Phn. No), Statistical extraction which is used to extract names, geographical places and phrase extraction. We can also use human aided process by adding templates and using it as reference.

2. Automatic Text Summarization: In natural language processing (NLP) automatic text summarization is one of the major problem which shows that how a computer can understand, analyse and derive meanings from a human language. Extracting of information from single or multiple documents is a very labourious , difficult and time taking task for a human being. So here comes the important role of automatic text summarization which solves this problem efficiently. The overall goal is to provide the useful information in a shortest possible compact size while preserving the original essence of the text.

## II. Types of Summarization

### A. Extractive and Abstractive Summarization

An extractive summarization method [1] begins with selecting important sentences, crucial paragraphs etc. from the source and remoulding them in a compact form. Whether a sentence is important or not , it's decision is based on many linguistic and statistical features leading the sentence. Through the method of scoring scheme the importance of a sentence is found out. Then the statements with higher score are selected to generate the summary. Rate of compression is a deciding factor in determining summary length. Whereas abstractive summarization produces an abstract summary including words and phrases that are different from those occurring in the original document. Hence abstract is a type of summary that involves ideas or concepts from the original document but are produced in different form. Therefore the concepts of natural language processing[2] are required here. Hence, it is harder to develop summaries than extractive summarization.

### B. Single document and Multiple document

When summary is made from a single source it is known as Single document type. But when summary is generated from more than one source document it is known as multiple document summarization. It is quite difficult as compared to single document because redundancy is increased in the process. Systems are there to check any redundant content. It first feeds the summary with selecting the sentences initially and later on matching already fed sentences with the newer ones coming, and if there is match these words/sentences are rejected otherwise it selects them[3]. Maximal Marginal Relevance[4] approach is used as suggested by Carbonell and Goldstein for reducing redundancy.

### C. Generic and Query focused

Summaries can also be of two types: generic or query-focused [5][6][7]. Topic-focused or user-focused summaries are same as query-focused summaries. In query focused summary the summary is generated on the basis of words or sentences required specifically by the user, whereas a general set of information is delivered in a generic summary.

### D. Supervised and Unsupervised

Summarization can be of supervised or unsupervised types also [8] . In supervised systems, there is initial important data known as training data which plays a major role in deciding the priority of important contents from the document. Hence there is huge requirement of large amount of labeled data. Sentence based classification comes into the picture. Those sentence which are present in the summary are called as positive samples and other being the negative samples [9] . Whereas in unsupervised no training data is used, in such heuristic rules are applied to extract highly relevant sentences to form a summary [10].

### E. On the basis of Language

Three types of summaries are there on the basis of language: mono-lingual, multi-lingual, and cross-lingual summaries. When source and target document are present in same language it is called mono-lingual summarization system. When the document whose summary is to be made contains a mixture of a number of languages like English, Hindi, Tamil, Telugu, Marathi etc. and the resulting summary also needed or generated in the same languages, then it is termed as a multi-lingual summarization system. If the original source document is in English and the summary generated is in any other language, then this type of summary is known as a cross-lingual summarization system.

Abstractive and Extractive summarizations are discussed in detail as given below:

## III. Abstractive Text Summarization Methods

It is divided into two categories namely Structured based and semantic based Approach

### A. Structured based Approach

It extracts most important information through the text. Different approaches such as tree based , rule based , template based are briefly discussed below.

*1) Tree based method:* Dependency tree[11] is used to make summaries. It can use language generator or an automatic summary generator. It takes into account units of given document and summarizes accordingly. It does not provide with a complete model to represent entire document.

*2) Rule based method:* In this summarized documents are presented as categories. Creates summaries with greater information density than original document. But it relies too much on human effort, so process can be slow and exhausting.

*3) Template based method:* It makes a template of the text. Linguistic patterns are used to map the document to template slots. It creates coherent summary because it relies on relevant information provided. Only disadvantage being construction of templates can be difficult and tedious.

*4) Ontology based method:* Knowledge base[12] is used for summarization. It employs fuzzy ontology logic to handle uncertain data. Handles uncertainty much better than previous models. Creating a rule based system for ontology is difficult task.

*5) Lead and Body Phrase Method:* This method uses insertion and substitution that have same head chunk in lead and body. Good for revision of a lead sentence which is semantically correct. But it consist of lot of repetition and focus is always on rewriting techniques.

### B. Semantic based approach

It deals with the linguistic data. It focuses on noun and verb phrases.

*1) Information item based Method:* It results is a well-defined information with minimal redundancy. It's inability in creating meaningful and grammatical sentences were the main cause of its failure. Very poor linguistic quality was seen in this method.

*2) Semantic graph based method:* It produces concise, coherent, less redundant and grammatically correct sentences. But only single document abstractive summarization is possible.

*3) Multimodal semantic model:* Produces abstract summary with excellent coverage because of its salient textual and graphical content. But can only be evaluated manually.

## IV. Extractive Text Summarization Methods

### A. Features

It focuses on extracting most important paragraphs and sentences. Both word level[24] and sentence level extraction are done in this summarization.
Word level features:

*1) Title word:* Words that appear in the title are important one to be included in the summary.

*2) Content word:* The important keyword such as noun, verb, adjective and adverbs which are perfectly good candidates to contribute to the summary.

*3) Biased word:* The domain specific words are predefined set of words which represent the theme of the document.

*4) Cue word:* The impact of the positivity or the negativity of the word in sentence formation is frequently of extreme importance. Example cues are like ”in conclusion” , ”in short” ,”the author says” etc.

*5) Uppercase and Quoted words:* The important words are mainly in uppercase letters and in single and double quotes also. So it becomes important to include those word in the final summary. Example : WTO, USA etc
Sentence level features:

*6) Length:* Length plays a major part in summary generation. Shorter sentences generally doesn't carry important information as compared to longer ones. The normalized length of the sentence is the ratio between the number of words occuring in the sentence to the number of words present in the maximum length sentence in the document.

*7) Location:* The important words , sentences are generally located at the start and end of the paragraph or the whole document. Hence in most of the summaries it was evident to include them as important information which should be included to make a better and meaningful summary.

### *B. Unsupervised Learning Methods*

These methods doesn't require user input in deciding the important points in the summary. Hence these provide better degree of automation in comparision to Supervised methods.

*1) Graph based approach:* Graphs can easily describe the information in any document.In this Extractive text summarization using external knowledge from Wikipedia incorporating bipartite graph framework [16 ]has been used. Concept similar to HITS algorithm were proposed by them which efficiently and coherently selects the important sentences. Eigen vector based approach know as LexRank[17]. The sentences represented as graph and the edges represent the similarity. The sentences were then clustered based on their LexRank Scores similar to PageRank algorithm[18].Advantages- Effective in areas such as image captions, biomedical documents and newswire, Coherency is improved, Redundant information is identified. Limitations- Doesn't focus on dangling anaphora issue.

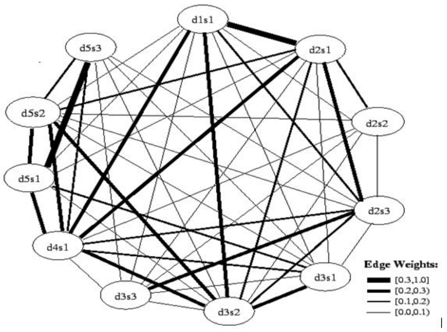

Fig 1: Weighted cosine similarity graph for the cluster

*2) Fuzzy logic:* As evident from diagram there are four basic components namely Fuzzifier, Inference Engine, Fuzzy Rule base and Deffuzifier. Source document is preprocessed with the features such as sentence length, sentence location, sentence similarity etc. To maintain coherency summary is generated in the order of occurances in the main document. Advantage-Coherency is improved. Limitation-Membership function and work of the fuzzy system.

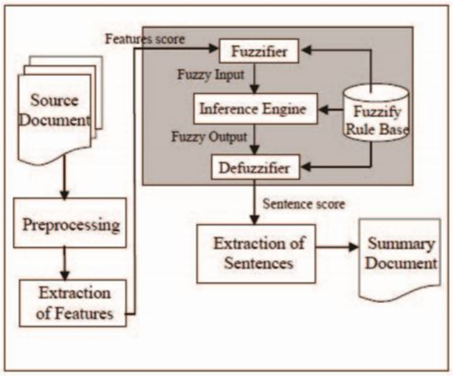

Fig 2: Fuzzy logic based text summarization

*3) Concept based Approach:* In this method concepts are extracted from external knowledge base such as Wikipedia. Importance of statements lies in the facts/concepts derived from wikipedia. The steps involved in this type of summarization can be describes as: (i)Retrieve concept of text from external knowledge base (ii) Build a graph model to depict the relationship between concepts derived and sentences (iii) generation of summaries based on the relative score of sentences.Advantage-Incorporation of similarity measures to reduce redundancy. Limitations-Dangling anaphora , verb references not considered.

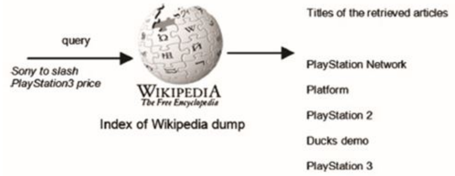

Fig 3: Concepts received for sentences from wikipedia[14]

*4) Latent Semantic Analysis Method(LSA):* External training is not required in this analysis. It takes as input the text of the document and search for patterns such as words that frequently occur together or words that are seen in different sentences[13]. If the common words are high in number that indicates that the sentences are semantically related. Singular Value Decomposition[19] method is used for finding these type of interrelations between words and sentences. Advantage: words and documents were mapped to the same concept space furthermore cluster of similar words were formed. Limitations: can't handle words with multiple meanings efficiently and assumes Gaussian distribution which is not suitable for all problems.

### *C. SUPERVISED LEARNING METHODS*

These methods include classification of sentences into summary generating sentences and non-summary generating sentences. Hence initial training data is required to train the model.

*1) Machine Learning Approach based on Bayes' Rule:* The machine is fed with training data to classify the input document into summary and non-summary generating sentences with the help Bayes Theorem. Advantage-Large set of training data improves sentence selection capability. Limitation-Human aid is required in initial stages of training.

*2) Neural network based Approach:* Neural networks are used to identify the important sentences from the document using RankNet Algorithm[20]. Two layer with back propagation was used. Firstly the . Training the machine using machine learning algorithm on test data. Three layer model can also be used. Important step involves the relationships establishments. (1) eliminating infrequent features (2) collapsing frequent features after which sentence ranking is done to identify the important summary sentences. Advantage-Customizable summaries according to user requirements. Limitations-Human aid is required, Slow in training and application, difficult to trace how machine makes the decision.

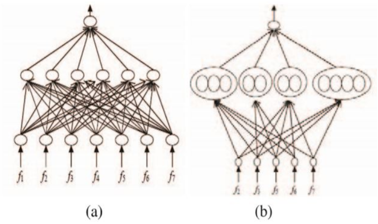
(a) (b)

Fig 4: (a)Neural network after training (b) after pruning

## V. TEXT SUMMARIZATION -INDIAN LANGUAGES

*1) Hindi Language:* Manjula Subramanyan et al. [21] presented a representation test model for Hindi text. They proposed an approach where the Hindi text document was fed and preprocessed followed by creating a semantic graph known as Rich Semantic Graph and then generating the reduced subgraph and finally creating an abstractive summary of the original document. The three phases are shown in the figure. The process starts with deep syntactic analysis of input text, then generates typed dependency relations (grammatical relations), and syntactic and morphological tags for each word [21]. Ontological domain to validate and couple the sentences for forming the rich semantic subgraphs which later on were merged with each other to define and precisely create the ultimate rich semantic graph.

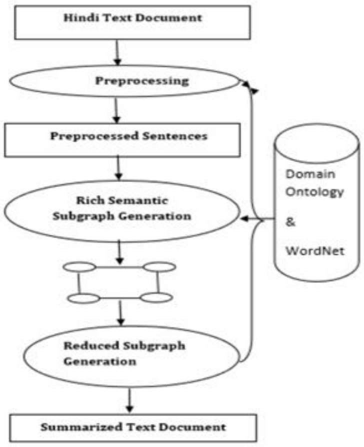
Hindi Text Document
Preprocessing
Preprocessed Sentences
Rich Semantic Subgraph Generation
Domain Ontology & WordNet
Reduced Subgraph Generation
Summarized Text Document

Fig 5:Architecture of Hindi summarization [21]

*2) Telugu Language:* Jagadish S. et al.[22] developed a model of abstractive text summarization method. Class based templates and IE rules were efficiently used. The F score was 0.815, precision being 0.8642 and Accuracy was 0.7217.

*3) Malyalam Language:* R. Kabeer et al.[23] proposed a graph and semantic based approaches to extract semantic triplets (Subject-Object-Predicate) from sentences in the document. A subgraph was selected using various Machine Learning techniques which falls in the classification domain. Naive Bayes method, Neural networks and Hidden Markov Model (HMM) are some of the machine learning approaches used for text summarization.

## VI. CONCLUSION

The research community is trying to develop more and more meaningful and coherent summaries where machine-generated summaries matches the capabilities of human-made summaries. Several works and techniques have been done to develop summaries till date. Although it is not possible to explain details of all all possible algorithms for text summarization in detail altogether in this paper, we tried to provide an insight into recent works and developments done in the field of automatic text summarization but we need a more reliable and clear-cut solution to produce useful, informative and well-organized summaries in a time-efficient manner which opens a whole new scope for further researches like integrating deep learning and neural networks techniques with the existing automatic summarization methods.

## REFERENCES

[1] Vishal Gupta and Gurpreet Singh Lehal, A Survey of Text Summarization Extractive Techniques, Journal of Emerging Technologies in web intelligence, volume 2, no.3, August 2010.

[2] Mahak Gambhir and Vishal Gupta, Recent automatic text summarization techniques: a survey, Springer Science+Business Media Dordrecht , 29 March 2016.

[3] Sarkar K (2010) Syntactic trimming of extracted sentences for improving extractive multi-document summarization. J Comput 2:177184

[4] Carbonell JG, Goldstein J (1998) The use of MMR, diversity-based re-ranking for re-ordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, pp 335336

[5] Gong Y, Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24st annual international ACM SIGIR conference on research and development in information retrieval. pp 1925

[6] Dunlavy DM , OLeary DP , Conroy JM, Schlesinger JD(2007) A system for querying, clustering and summarizing documents. Inf Process Manag 43:15881605

[7] Ouyang Y, Li W, Li S, Lu Q (2011) Applying regression models to query-focused multi-document summarization. Inf Process Manag 47:227237

[8] Riedhammer K, Favre B, Hakkani-Tur D (2010) Long story short-global unsupervised models for keyphrase based meeting summarization. Speech Commun 52:801815

[9] Song W, Choi LC, Park SC, Ding XF (2011) Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization. Expert Syst Appl 38:91129121

[10] Fattah MA, Ren F (2009) GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Comput Speech Lang 23:126144. doi:10.1016/j.csl.2008.04.002

[11] 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.

[12] 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.

[13] 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 and Knowledge Management Process (IJDKP), 2013, Vol.3, No.1.

[14] 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

[15] Khan Atif, Salim Naomie, A review on abstractive summarization Methods, Journal of Theoretical and Applied Information Technology, 2014, Vol. 59 No. 1.

[16] Y. Sankarasubramaniam, K. Ramanathan, and S. Ghosh, Text summarization using wikipedia, Information Processing Management, vol. 50, no. 3, pp. 443-461, 2014.

[17] G. Erkan and D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, Journal of Articial Intelligence Research, pp. 457-479, 2004

[18] S. M. R .. W. T. L., Brin, The pagerank citation ranking: Bringing ordertotheweb,Technicalreport,StanfordUniversity,Stanford,CA., Tech. Rep., (1998).

[19] M. G. Ozsoy, F. N. Alpaslan, and 1. Cicekli, Text summarization using latent semantic analysis, Journal of Information Science, vol. 37, no. 4, pp. 405-417, 2011.

[20] K. M. Svore, L. Vanderwende, and C. J. Burges, Enhancing single document summarization by combining ranknet and third-party sources. in EMNLP-CoNLL, 2007, pp. 448-457.

[21] Manjula Subramaniam, Prof. Vipul Dalal, Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method, (IRJET), vol. 02, no. 02, May 2015.

[22] Jagadish S. Kallimani, K. G. Srinivasa and B. Eswara Reddy, Statistical and Analytical Study of Guided Abstractive Text Summarization Current Science, 2016, Vol. 110 No. 1.

[23] R. Kabeer, M I Sumam, Text Summarization of Malayalam Documents-an Experience International Conference on Data Science and Engineering(ICDSE), 2014.

[24] Mathews, Lincy Meera, and E. Sathiyamoorthy. ”Intricacies of an Automatic Text Summarizer.” International Journal of Engineering and Technology (IJET) 5.3 (2013).