# Extractive Approach for Text Summarization Using Natural Language Processing

**Vijayalakshami K[1], Trupti Patil[2], Kajal Shele[3], Dr. D. R. Ingle[4]**

*[1]Vijayalakshami K, Dept, of Computer Engineering, Bharati Vidyapeeth College of Engineering, Maharashtra, INDIA*
*[2]Trupti Patil, Dept, of Computer Engineering, Bharati Vidyapeeth College of Engineering, Maharashtra, INDIA*
*[3] Kajal Shele, Dept, of Computer Engineering, Bharati Vidyapeeth College of Engineering, Maharashtra, INDIA*
*[4]Dr. D. R. Ingle, Head of Deptartment of Computer Engineering, Bharati Vidyapeeth College Of Engineering, Maharashtra, INDIA*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *There is a vast amount of textual content, and it is only growing every single day. Think of the Internet, including web pages, news articles, status updates, blogs, and more. The data is unstructured and the best we can do is use search to navigate it and skim the results. There is a lot of need to reduce this text data, focusing summaries that capture key details, both so we can navigate it more effectively as well as check whether large documents Contains the information the user is looking for. The natural language processing research community is developing new ways to summarize the text. The Extractive Approach to Text Summarization is a process of summarizing an original text that contains important information. Text summarization is a process of preparing summaries by reducing the size of the original document and retaining important information of the original document. The paper explains in detail two main categories of text summarization methods. These are extractors and abstract summarization methods.*

***Key Words***: **NLP, Summarization, Text, Abstractive Summary**

## 1. INTRODUCTION

Natural Language Processing is a domain of AI, computing and linguistics focused on the interaction between computers and human language. Natural language processing may be a process of developing a system which will process and produce language nearly as good as human can produce. The use of World Wide Web has increased then the matter of data overload also has increased. Hence there's a requirement of a system that automatically retrieves, categorize and summarize the document as per users need. Document summarization is one possible solution to this problem.

More and more electronic data is available on the Internet and it is not possible to read everything and hence some form of information condensation is needed. Summarization serves as a tool which helps the user to efficiently find useful information from immense amount of information.

Text summarization can be used by various applications for instance researchers need a tool to generate summaries. For deciding whether to read the entire document or not and for summarizing information searched by user on Internet. News groups may utilize the multi document summarization technique to cluster the information from different media and summarize. A summary is basically a text that is produced from one or more texts, that provides only the important information in the original text, and it is of a shorter form. The objective of text summarization is transforming the source text into a shorter version with semantics. The most important benefit of using a summary is, it saves the reading time.

## 2. LITERATURE REVIEW

Past literature that utilizes the various summarization methods are cited in this section. Most of the researchers concentrate on sentence extraction instead of generation of the text for text summarization. The most widely used technique for summarization is based on statistical features of the sentence which gives extractive summaries.

Luhn proposed that the most frequent words represent the most important concept of the text. Luhn's idea was to give the score to each sentence based on the number of occurrences of the words and then select the sentence which is having the highest score. Edmunson proposed methods based on title, location and cue words. He stated that first few sentences of a document or first paragraph contains the information of the topic and that should be included in the summary. One of the limitation of statistical approach is that, they do not consider semantic relationship among sentences. Goldstein proposed a query-based summarization to generate a summary by extracting relevant sentences from a document based on the query fired. The criterion for extraction is provided as a query. The probability of being included in a summary increases in accordance to the number of words co-occurred in the query and a sentence. Goldstein also studied news article summarization and used statistical and linguistic features to rank sentences in the document.

One of the summarization can be implemented by sentence extraction and clustering. As per LI Cun & ZHANG Pei-ying suggestions, sentences are clustered based on the semantic distance between sentences and then calculates the accumulative sentence similarity between the clusters and finally select the sentences on the basis of the extraction rules. K-means algorithm is used to cluster sentences.

The concept of lexical chain was first proposed by Hirst and Morris. Lexical chains take advantage of the cohesion among an arbitrary number of related words. Lexical chains are created by grouping set of words that are semantically related. Barzilay and Elhadad constructed lexical chain by calculating semantic distance between words using WordNet. Strong lexical chains are chosen and the sentences related to these strong chains are selected as a summary.

H. Gregory Silber and McCoy developed a liner time algorithm for lexical chain computation. The paper discusses an algorithm for creating lexical chain which creates an array of Meta-Chain whose size is the number of nouns senses in the Word Net and in the document. There were some problems with the algorithm like proper nouns and anaphora resolution that were to be addressed.

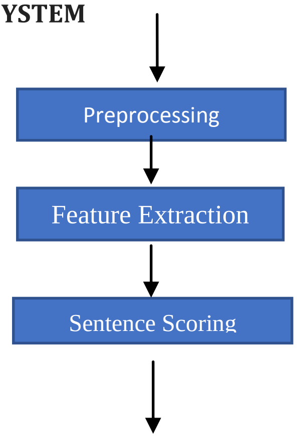
There is another method for summarization by using graph theory. The author discovered a method based on subject-object-predicate (SOP) triples from individual sentences to create a semantic graph of the original document. The relevant concepts, carrying the meaning, are split across clauses. The author suggested that identifying and exploiting links among them could be useful for extracting relevant text.

One of the researchers, Pushpak Bhattacharyya from IIT Bombay introduced a Word Net based approach for summarization. By generating a sub-graph from Word-net,document are summarized. With respect to the synsnet weight are assigned to nodes of the sub-graph using the Word Net. Either statistical approach or linguistic approach or a combination of both are use by most common text summarization techniques.

## 3. ALGORITHM

A. Collect data
B. Clean up data
C. Algorithms NLTK to build tokens (word or sentences)
D. Word frequency
E. Weighted frequency for each words
F. Calculate score for each sentences
G. Select top sentences for summary

## 4. PROPOSED SYSTEM



Text Summarization

Text summarization approach consists of following stages:

A. Preprocessing

B. Feature Extraction

C. Sentence Scoring

A. Text Preprocessing

There are four steps in preprocessing:

1. Segmentation: Document are divided into sentences.

2. Removal of Stop words: Stop words are frequently occurring words such as 'a' an', the' that provides less meaning and contains noise. The Stop words are stored in an array and those are predefined.

3. Word Stemming: converts every word into its root form by removing its prefix and suffix so that it can be used for comparison with other words.

B. Feature Extraction:

The text document is represented by set, D= {S1, S2, - - - , Sk} where, Si denotes a sentence contained in the document D. The document is belong to feature extraction. The important word and sentence features to be used are decide. Title word, Sentence length, Sentence position, numerical data, Term weight, sentence similarity, existence of Thematic words and proper Nouns this features are use by feature extraction.

C. Sentence Scoring:

Each sentence is evaluated by considering the linear combination of multiple parameters like frequency, sentence position, cue words, similarity with title, sentence length and proper noun. With respect to the scores sentences are rank.

## 5. CONCLUSIONS

Need to develop efficient and accurate summarization systems. Lots of research still going on this field especially in evaluation techniques. Extractive techniques are usually use rather than abstractive techniques. Our algorithm will shows better results as compared to the output produced by online summarizers. There is information overload due to rapid growth of technology and use of Internet. Strong text summarizers can solved this problem, which produces a summary of document that help user. That's why there is a need to develop system where a user can efficiently get a summarized document. One possible solution is to use either extractive or abstractive methods. Text summarization by extractive is easier to build.

## 6. REFERENCES

[1] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh.A Comprehensive Survey on Text Summarization Systems. 2009 In proceeding of: Computer Science and its Applications, 2nd International Conference.

[2] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999.Summarizing text documents: Sentence selection and evaluation metrics.

[3] Horacio, L. Guy.Generating indicative-informative summaries with SumUM: Summarization. Computational linguistics - Association for Computational Linguistics, 2002.

[4] Luhn, H.P., 1959. The automatic creation of literature abstracts.

[5] ZHANG Pei-ying, LI Cun-he. Automatic text summarization based on sentences clustering and extraction.

[6] Barzilay, R., Elhadad, M.Using Lexical Chains for Text Summarization. Workshop on Intelligent Scalable Text summarization, Madrid, Spain,1997.

[8] Eduard Hovy and Chin Yew Lin.Automated text summarization in 1999.

[9] Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics.

[10] Silber G.H., Kathleen F. McCoy. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. Computational.

[11] Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya.Generic Text Summarization Using Word net. Language Resources Engineering Conference.

[12] J. Leskovec, M. Grobelnik, N. Milic-Frayling.Extracting Summary Sentences Based on the Document Semantic Graph. Microsoft Research, 2005.

[13] D. Radev, E. Hovy, K. McKeown, "Introduction to the Special Issue on Summarization", Computational Linguistics.