# Content Summarization System

**Minal S. Khodape[1], Rohini R. Patil[2], Nayana G. Gawande[3], Bhagyashri N. Vanjari[4], Suvarna M. Borse[5]**
[1, 2, 3, 4, 5] Dept of Computer Engineering
[1, 2, 3, 4, 5] SSBT College of Engineering and Technology Bambhori, Jalgaon

***Abstract-*** *The document that published in chronological sequence. Topic is the sentence which captures the attention of the readers, and the core parts of the topic will be associated temporarily so that they would help the document readers to grasp its content easily. Effective automatic summarization has become more important for the increase in the amount of content available online, fast. We use the extraction based techniques to generate automatic summaries from a huge document. Automatically creating a short version of a given multiple text that provides useful information for the user is called text summarization. Summarization methods depends on covering wider area of information which comprises of important information. Topic-Oriented summaries focus on interest of user's interested topics and extract the information that is related to specified topic. There are basically two methods of summarization: extraction based and abstraction based..*

***Keywords***- Text mining, Language summarization, Text analysis, Preprocessing, WordNet.

## I. INTRODUCTION

There are many other means to get information rather than traditional media. Internet documents play the vital role here. Everything became computerized, and even the day to-days news. News that is provided in the internet are more in number and it is Highly impossible for the readers to read through all the related documents. According to WordNet summary is defined as a brief statement that presents the main Points in a concise form. Summarization process involves interpretation, transformation and generation. There are two types of summarization is present first is abstraction-based summarization and second is extraction -based summarization . In abstract-based summarization, an abstract is created by interpreting the text contained in the original document and generating summary that express the same in a more concise way. In extractive-based approach form a meaningful summary is generated from the original text . In this approach sentences are given scores based on different feature and sentences with higher rating are selected for summary. It uses various Natural Language processing approaches for information retrieval. Summarization process can also be classified based upon the number of source documents, task specific constraints and use of external resources. Summarization is classified as single-document or multi-document based upon the number of source document. In multi document summarization information overlaps between different document makes task difficult. Based upon external resources summarization can also be classified as knowledge- rich or knowledge-poor. Knowledge rich summarizer uses external source external corpus like Wikipedia, Word Net etc. In query-focused or query oriented summarization summary is constructed with information relevant to the query.

## II. LITERATURE SURVEY

**Abdel Fattah** : Abdel Fattah, Mohamed has proposed a simple approach for text summarization. They have considered features like position, name entities, numerical data, length, vocabulary overlaps etc. to generate summary.

**Ryang , D. Seonggi** : Ryang, D. Seonggi proposed a method of automatic text summarrization with reinforcement learning. Researches have also been done for summarization of Wikipedia articles.

**Kamal Sarkar** : Kamal Sarkar is built for summarization of medical documents using machine learning approach.

## III.PROPOSED SYSTEM

In this section, we mention preprocessing methods used in the proposed content summarization system

A:Document: The first preprocessing step is the selection of document in the form of text.

B: Sentence: In the second preprocessing step, split the document by dot (.). Also give the number of sentences in the document.

C: Stopwords: In the third preprocessing step, remove less important words like an, the, on, in, a.

D: Unique Words: In Unique Words find the unique words and give count to each words. All unique words also called keywords.

E:Stemming: In this, create the baseword using the porter stemmer algorithm. The base word is taken.e.g.Complete is taken instead of completing, completed.

Porter Stemmer Algorithm:

Content Summarization System basically has porter stemmer algorithm for proper functioning:-

The porter stemming algorithm is a process of removing suffixes from words in English. Removingsuffixes automatically is an operation which is especially useful in the field of information
retrieval.

1. Replace sses by ss (eg.caresses- caress).
2. delete s if the precedding word part contains a vowel not immediately before the s.(eg. gaps-gap but gas )
3. Replace ied or ies by i if the precceded by more than one letter, otherwise by ie (ties-tie, ponies- poni)
4. If suffix is us or ss do nothing (eg.stress-stress).
5. If the suffix is s then removed it (eg cats-cat).
6. Replaced eed, eedly by ee if it is in the part of word after the first non vowel following vowel (eg. agreed-agree, feed-feed).
7. if the suffix is ing or ed then removed it (walking-walk, plasterd-plaster)

F: Significant: Count the frequency occurrences of each word for every document.

G: Weight: Weight is the preprocessing step in which term weight is given to words.

H: Ranking: Ranking is the preprocessing step in which after scoring of each sentence, sentences are arranged in descending order of their score value i.e. the sentence whose score value is highest
is in top position and the sentence whose score value is lowest is in bottom position.

I: Summary: Summary is the preprocessing step in which after the sentences based on their total score the summary is produced selecting X number of top ranked sentences where the value of X is provided by the user. For the readers convenience, the selected sentences in the summary are reordered according to their original positions in the document.

## IV. RESULT AND DISCUSSION

The content Summarization system basically first select the document then load the document and count the every sentences after dot(.). Stopword are removed from that document. After removed Stopword count of unique words means how many times that word is repeated. Then next step is stemming. Then Significant and term weight is generate of words. And then giving the ranking from low to high and finally Summary is generated for that document.

By applying all the preprocessing steps and porter stemmer algorithm it gives the result in the form of meaningful short summary. It gives the better result than the abstraction based summary.

## V. FEATURES

1. Summarize finds the main points and key details.
2. Saves time. It is better than manual summary.
3. Provide meaningful sentence to end users without changing meaning of original word.
4. Good readability.
5. No semantic redundancy.

## VI. CONCLUSION AND FUTURE SCOPE

An automatic Content text summarization approach by sentence extraction using an un-supervised learning algorithm is proposed. In contrast to supervised methods, does not need large amount of golden samples for training. Therefore, our project is more independent from language and domain. Here is an massive need for automatic summarization tools in this age of information overload. We emphasized various extractive approaches for single. Although it is not feasible to explain all diverse algorithms and approaches comprehensively , we think it provides a good insight into recent trends and progresses in automatic summarization methods.

One of the future plans may be to apply the topic-focused summarization framework to news articles or blogs and to extend the work in the machine leaning approaches. Topic- focused summaries of news articles would be lot more accurate and valuable to users. It would be more interesting to work on topic related information and summarization in the domain of social media in future and also as online summarizer or web page summarizer. The rate at which the information is growing is tremendous. Hence it is very important to build a multilingual summarization system. The work presented by the thesis can also be applicable to multi document summarization by using minimal extensions.

## VII. ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. Our sincere thanks to the Principal Prof. Dr. Kishor S. Wani, Shram Sadhana Bombay Trusts College of Engineering Technology, for providing us facilities and resources to complete project work. We also take this opportunity to thank all our friends for offering all possible help.

## REFERENCES

[1] A.Pearline Divya, S.Leela, The Content Summarization system", International Journal Of Advanced Research in Computer Engineering & Technology (IJARCET), 2013.

[2] Ayush Agrawal , Utsav Gupta, Extraction based approach for text summarization using k-means clustering", International Journal of Scientific and Research Publications, 2014.

[3] Deepali K. Gaikwad and C. Namrata Mahender, A Review Paper on Text Summarrization", International Journal of Advanced Research in Computer and Communication Engineering, 2016.

[4] G.V. Garje, PhD S.V. Khaladkar A. N. Khengare J.M. Pawar M.S. Vidhate, Genearting
Multi-Document Summarization using Data merging Technique", International Journal of computer Applications(0975-8887), 2016.

[5] Dimitrios Galanis, Automatic generation of Natural Language summaries", "Department of informatics Athens University of Economics and business", 2012.
{}[]A201