Abstract: Text summarization is a process of extracting or collecting important information from original text and presents that information in the form of summary. Text summarization has become the necessity of many applications for example search engine, business analysis, market review. Summarization helps to gain required information in less time. This paper is an attempt to summarize and present the view of text summarization from every aspect from its beginning till date. The two major approaches i.e., extractive and abstractive summarization is discussed in detail. The technique deployed for summarization ranges from structured to linguistic. In Indian many languages also the work has been done, but presently they are in infancy state. This paper provides an abstract view of the present scenario of research work for text summarization.

Keywords: Text Summarization, Natural Language Processing, Extractive Summary, Abstractive Summary.

I. INTRODUCTION

Text summarization is a process of extracting or collecting important information from original text and presents that information in the form of summary. In recent years, need for summarization can be seen in various purpose and in many domain such as news articles summary, email summary, short message of news on mobile, and information summary for businessman, government officials, researchers online search through search engine to receive the summary of relevant pages found, medical field for tracking patient’s medical history for further treatment [1].

On the internet, many such examples are available like, news article summarizer such as Microsoft News2, Google1 or Columbia Newsblaster3 [1]. BaseLine, FreqDist, SumBasic, MEAD, AutoSummarize & SWESUM is few popular biomedical summarization tools [2]. Text Comapcter, Simplify, Tools4Noobs, FreeSummarizer, WikiSummarizer & SummarizeTool are online summarization tools. Open Text summarizer, Classifier4J, NCClassifier, CNGLSummarizer are few widely used open source summarization tools [3]. The need of having information in abstract form on a click has increased as, the need for automatic text summarization has also increased in many areas namely, news articles summary, email summary, short message news on mobile and information summary for business, government officials research, online research engines to receive summary. In late 1950, the first system came in term; The automatic summarizer in general selects important sentences from the document and groups them together, it consume less time or time saving to understand the content within the large document. The aim of automatic text summarization is to convert large document into shorter one and store important content [4]. 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 can be broadly divided into two groups: extractive summarization and abstractive summarization. Extractive summarizations extract important sentences or phrases from the original documents and group them to produce a summary without changing the original text. An extractive text summarization system is proposed based on POS tagging by considering Hidden Markov Model using corpus to extract important phrases to build as a summary [5]. Abstractive summarization consists of understanding the source text by using linguistic method to interpret and examine the text. Abstractive methods need a deeper analysis of the text. These methods have the ability to generate new sentences, which improves the focus of a summary, reduce its redundancy and keeps a good compression rate [1].

II. TEXT SUMMARIZATION FEATURES

Text summarizers identify and extract key sentences from the source text and concatenate them to form a concise summary. A list of features as discussed below can be used for selection of key sentences in Table 1 [2, 4].

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term Frequency</td>
<td>Salient terms provided by statistics are based on term frequency, thus salient sentences are those words that occur repeatedly [4]. The frequently occurring word increases score of sentences. The most common measure widely used to calculate the word frequency is TF IDF [2].</td>
</tr>
<tr>
<td>Location</td>
<td>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 [4]. First and last sentence of paragraph has greater chance to be included in summary [2].</td>
</tr>
<tr>
<td>Cue Method</td>
<td>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” [2].</td>
</tr>
</tbody>
</table>
III. TECHNIQUES USED FOR TEXT SUMMARIZATION

Text summarization as discussed is broadly divided into abstractive and extractive. The brief description about each approach is discussed in following section:

A. Abstractive Summarization Approach

Summarizations using abstractive techniques are broadly classified into two categories: Structured based approach and Semantic based approach [1].

1) Structured Based Approach:

Structured based approach encodes most important information from the document through cognitive schemes such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure [1].

Brief abstract of all the techniques under structured based approach is provided in Table 2.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Description</th>
<th>Advantages</th>
<th>Limitation</th>
<th>Author &amp; Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Based Method</td>
<td>- It uses a dependency tree to represent the text of a document.</td>
<td>- It walks on units of the given document read and easy to summary.</td>
<td>- It lacks a complete model which would include an abstract representation for content selection.</td>
<td>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].</td>
</tr>
<tr>
<td>Template Based Method</td>
<td>- It uses a template to represent a whole document.</td>
<td>- It generates summary is highly coherent because it relies on relevant information identified by IE system.</td>
<td>- Requires designing of templates and generalization of template is to difficult.</td>
<td>Harabagiu and Lacatusu (2002) [2], Tatsuro Oya, Yashar Mehdad, Giuseppe Carenni, Raymond Ng (2014) [19].</td>
</tr>
<tr>
<td>Ontology Based Method</td>
<td>- Use ontology (knowledge base) to improve the process of summarization.</td>
<td>- Drawing relation or context is easy due to ontology</td>
<td>- This approach is limited to Chinese news only.</td>
<td>Lee and Jian (2005) [2], Meghana viswanath(2006) [9], Ramezani Majid, Feizi-Derakhshi Mohammad-Reza(2015) [20], R. Ragunath and N. Sivaranjani (2015) [23].</td>
</tr>
<tr>
<td>Lead and Body Phrase Method</td>
<td>- 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</td>
<td>- It is good for semantically appropriate revisions for revising a lead sentence.</td>
<td>- Parsing errors degrade sentential completeness such as grammaticality and repetition.</td>
<td>Tanaka and Kinoshita (2009) [2].</td>
</tr>
</tbody>
</table>
order to rewrite the lead sentence. lacks a complete model which would include an abstract representation for content selection.

- Documents to be summarized are represented in terms of categories and a list of aspects.
- It has a potential for creating summaries with greater information density than current state of art.
- The drawback of this methodology is that all the rules and pattern are manually written, which is tedious & time consuming.

Genest and Lapalme (2012)[2].

2) Semantic Based Approach

In Semantic based approach, semantic representation of document is used to feed into natural language generation (NLG) system. This method focuses on identifying noun phrase and verb phrase by processing linguistic data [1]. Brief abstract of all the techniques under semantic based approach is provided in Table 3.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Description</th>
<th>Advantages</th>
<th>Limitation</th>
<th>Author &amp; Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multimodal semantic model</td>
<td>- A semantic model, which captures concepts and relationship among concepts, is built to represent the contents of multimodal documents.</td>
<td>- 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.</td>
<td>- The limitation of this framework is that it is manually evaluated by humans.</td>
<td>Greenbacker (2011)[2].</td>
</tr>
<tr>
<td>Information Item Based Method</td>
<td>- 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.</td>
<td>- The major strength of this approach is that it produces short, coherent, information rich and less redundant summary.</td>
<td>- It rejected due to the difficulty of creating meaningful and grammatical sentences from them. - Linguistic quality of summaries is very low due to incorrect parses.</td>
<td>Genest and Lapalme (2011)[2], Daniel Mallett, James Elding, Mario A. Nascimento (2004)[15].</td>
</tr>
<tr>
<td>Semantic Graph Based Method</td>
<td>- 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.</td>
<td>- It produces concise, coherent and less redundant and grammatically correct sentences.</td>
<td>- This method is limited to single document abstractive summarization.</td>
<td>Moawad &amp; Aref (2012)[2], Kavita Ganesan, ChengXiang Zhai &amp; Jiawei Han, (2010)[30], Laura Plaza, Alberto Díaz &amp; Pablo Gervás, (2011)[31], Manjula Subramaniam, Prof. Vipul Dalal(2015)[37].</td>
</tr>
</tbody>
</table>

B. Extractive Summarization Techniques

An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences [4]. Table 4 shows the brief abstract of techniques of extractive based approach.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Description</th>
<th>Author &amp; Year</th>
</tr>
</thead>
</table>
| Term Frequency- Inverse Document Frequency Method | -Sentence frequency is defined as the number of sentences in the document that contain that term.  
-Then this sentence vectors are scored by similarity to the query and the highest scoring sentences are picked to be part of the summary. | M.Fachrurozi, Novi Yusliani, and Rizky Utami Yoanita, (2013) [27], Mr. S. Pimpalshende A. N. (2013) [38]. |
| Cluster Based Method                   | -It is intuitive to think that summaries should address different “themes” appearing in the documents.  
-If the document collection for which summary is being produced is of totally different topics, document clustering becomes almost essential to generate a meaningful summary.  
-Sentence selection is based on similarity of the sentences to the theme of the cluster (Ci).  
The next factor that is location of the sentence in the document (Li). The last factor is its similarity to the first sentence in the document to which it belongs (Fi).  
\[ S_i = W_1 * C_i + W_2 * F_i + W_3 * Li \]  
Where, W1, W2, W3 are weight age for inclusion in summary.  
| Graph Theoretic Approach               | -Graph theoretic representation of passages provides a method of identification of themes.  
-After the common pre-processing steps, namely, stemming and stop word removal; sentences in the documents are represented as nodes in an undirected graph. | Rada Mihalce, Niraj Kumar, Kannan Srinathan and Vasudeva Varma, (2013)[33], Sarda A.T. and Kulkarni A.R.(2015)[34]. |
| Machine Learning Approach             | -The summarization process is modelled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess.  
The Classification probabilities are studied statistically using Navie Bayes Classifier rule:  
\[ P(\text{s} \in S \mid F_1, F_2, ..., F_N) = \frac{P(F_1, F_2, ..., F_N \mid \text{s} \in S) \ast P(\text{s} \in S)}{P(F_1, F_2, ..., F_N)} \] | Kamal Sarkar, Mita Nasispuri, Suranjian Ghose(2011)[10], Mehnnoosh Bazrfigan And Muosa Radmanesh(2014)[35]. |
| LSA Method                             | -It gets this name LSA because SVD applied to document word matrices, group documents that are semantically related to each other, even when they do not share common words. | Patil Pallavi D, Man P M(2014)[22], Hanane Froud, Abdelmonaine Lachkar and Said Alaoui Ouatik (2013)[23], Mr. S.A.Babar and Prof. S.A. Thorat(2014)[28], Josef Steinberger, Karel Jezek (2014)[36], Otsoyo Makbule Gulcin, Cicekli Ilyas and Alpaslan Ferda Nur (2010)[40]. |
| Text summarization With Neural Networks | -This method involves training the neural networks to learn the types of sentences that should be included in the summary.  
| Automatic TS based on fuzzy logic      | -This method considers each characteristic of a text such as similarity to title, sentence length and similarity to key word etc. as the input of the fuzzy system. | Ladda Suamnali, Naomie Salim, and Mohammed Salem Binwahlan (2009)[11], Ms. Pallavi D.Patil, Prof. N.J. Kulkarni (2014) [22], Patil Pallavi D., Man P M(2014) [25], Rucha S. Dixit, Prof. Dr. S. S. Apte(2012)[26], Mr. S.A. Babar and Prof. S.A. Thorat(2014) [28], S. Santha Megala Dr. A. Kavitha Dr. A. Marimuthu (2014)[29]. |
| Query Based Extractive Text Summarization | -In query based text summarization system, the sentences in a given document are scored based on the frequency counts of terms.  

TABLE 4 EXTRACTIVE TEXT SUMMARIZATION TECHNIQUES [4]
### IV. TEXT SUMMARIZERS FOR INDIAN LANGUAGES AND COMPARISON OF THEIR PERFORMANCE

#### TABLE 5
PERFORMANCE COMPARISON OF EXISTING INDIAN SUMMARIZERS

<table>
<thead>
<tr>
<th>Language</th>
<th>Researchers</th>
<th>Method</th>
<th>Features /Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dr. Latesh Malik (2013)</td>
<td>Single document Summarization using Extraction Method.</td>
<td>It uses statistical &amp; linguistic feature &amp; also uses Genetic Algorithm [41].</td>
</tr>
<tr>
<td></td>
<td>Anitha J., Prof. P. V. G. D. Prasad Reddy and M. S. Prasad Babu (2014)</td>
<td>Text Summarization using Extractive Method.</td>
<td>It uses fuzzy classifier and Neural Network. (Precision = 0.90, Recall = 0.88) [43].</td>
</tr>
<tr>
<td>Punjabi Text</td>
<td>Vishal Gupta (2010)</td>
<td>News Document using Extractive Method.</td>
<td>-Accuracy of the system is varies from 81% to 92 % [41]. -It consist of two phases: i)Pre-processing, ii) Processing [38].</td>
</tr>
<tr>
<td>Tamil Text</td>
<td>Gupta (2012)</td>
<td>Text Summarization using Extractive Method.</td>
<td>It uses TF-IDF techniques. (F-scores 97.87%, 95.32% &amp; 94.63%) [6].</td>
</tr>
<tr>
<td></td>
<td>Kumar and Devi (2011)</td>
<td>Text Summarization using Extractive Method.</td>
<td>It uses graph theoretic scoring technique [6].</td>
</tr>
<tr>
<td></td>
<td>Das and Bandyopadhyay (2010)</td>
<td>Text Summarization using Extractive Method.</td>
<td>It uses k-means approach And Page Rank standard approach. Precision= 72.15%, Recall=67.32% and F-Score= 69.65 % [6].</td>
</tr>
<tr>
<td></td>
<td>Jagadish S, Kallimani, K. G. Srinivasa and B. E. Reddy (2016)</td>
<td>Text Summarization using Abstractive Method.</td>
<td>It uses IE rules and class based templates. (F-score- 0.815, Precision-0.8642, Recall-0.7973, Accuracy- 0.7217) [49].</td>
</tr>
</tbody>
</table>
V. CONCLUSION

Text summarization 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. Precise information helps to search more effectively and efficiently. Thus text summarization is need and used by business analyst, marketing executive, development, researchers, government organizations, students and teachers also. It is seen that executive requires summarization so that in a limited time required information can be processed. This paper takes into all about the details of both the extractive and abstractive approaches along with the techniques used, its performance achieved, along with advantages and disadvantages of each approach. 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. Through the study it is also observed that very less work is done using abstractive methods on Indian languages, there is a lot of scope for exploring such methods for more appropriate summarization.

REFERENCES

33. Sivakumar A. P., Premchand P, and Govardhan A., “Query-Based Summarizer Based on Similarity of Sentences and Word


