Survey on Summarization Techniques and Existing Work

Prof. Satish Kamble1, Shivlila Mandage1, Shubhangi Topale1, Dipali Vagare1, Prerana Babbar1,

1.1 Professor, PVG’s College of Engineering and Technology, Pune, Maharashtra, India
satishkamble53@gmail.com

1 Student, PVG’s College of Engineering and Technology, Pune, Maharashtra, India
{shivlilamandge32,shubhangitopale25,dipali.vagare1995,preranababbars@gmail.com

Abstract

The text document summarization is the concept of summarizing the text where the whole document is condensed to a smaller version retaining its original abstract meaning. In extractive summarization technique it selects the most important sentences from documents to make a summary. Due to increasing amount of news articles on web, it is very difficult for a user to read all the newspapers and to sort and extract all different perspectives of the news. In this survey paper we explore the most prominent and significant work done in the single and multiple document summarization on English as well as other regional languages (Marathi, Bangla). Whereas, for English language there are many techniques have been proposed, rare works has been done for Marathi text summarization. Usually summarization technique has two classes as Abstractive summaries and Extractive summaries.

Keywords: Text document summarization, Abstractive summaries, Extractive summaries;
1 Introduction

In this growing digital age, for supporting and interpreting text information text summarization has become a crucial and timely tool. Automatic text summarization is the methodology in which text is reduced to create summary from an original document. Text-summarization is an important application, given the exponential rise in data available on Net. Reduction of text is a complex problem which poses many challenges to the scientific community.

The main intention of the text summarization is to convey the important ideas in the documents, by eliminating less important and redundant pieces of information. About 71 million people speak Marathi as their native language. Marathi is one of the top 22 official languages of India. In recent years many efforts are undertaken to get Marathi data sources into main stream.

While summarization two fundamental questions arise:

- How to select essential content?
- How to express the selected content in a concise manner?

For summarization of English document many systems are available which provides adequate accuracy. Yet there is no such a system for summarization of Marathi document. This can be helpful for individuals who are trying to read and access large amounts of Marathi information in a short interval of time. The main aim of summarization is to summarize a document by

To extract important facts from document.
To cover its all silent points and aspects.
To narrow down on particular statement.
To include most important statements without any error or repetitions.

There are three sections of the summarization system: Pre-processing: the Text document step includes finding boundary of Marathi sentences, Tokenizing the sentence, removal of Marathi-stop-words, Stemming and breaking Marathi document into collection of sentences & Elimination of duplicate sentences. In sentence ranking: step, the importance of sentences are determined and calculated by considering different features. For selecting important sentences in summary distinct features are: Frequency of word, Positional value, cue-words, Skeleton of the Document (words in title and header) [38]. Summary: After ranking the sentence based on their score and selecting top ranked sentences the summary is formed. To increase the readability of summary, the sentences in summary is reordered based on their appearances in original text; for example, the sentence which occurs first in original text will appear first in summary.
2 Related Work

There are distinct automatic text summarization systems available for most of the commonly used natural language. For English text following work has been done in recent years.

The Luhn developed "The automatic creation of literature abstracts" in 1958 [2] which perform the summarization of a single-document and propose the frequency of a singular word in a document to be a valuable measure of significance. Though the technology developed by Luhn’s was a preliminary step towards the summarization, but his many ideas are still found to be a very effective for text summarization. Jing presented "Sentence Reduction for Automatic Text Summarization" in 2000 [3] the system which reduce a sentence by removing unsuitable phrases like prepositional phrases, clauses, to infinitives, or gerunds from sentences Edmundson presented "New methods in automatic extracting" 1969[4] and proposed a ordinary structure of text summarization methodology. From his previous work he had taken two ideas such as positional value and frequency of word. There were two primary features used the first one was skeleton of the document (the sentence containing a title or heading) and the second one was presence of cue words (occurrence of words like significant, or hardly).

In recent years, immense amount of text documents in Indian languages are up for grabs on internet. For better management and retrieval of such documents, automatic classification can be helpful. Till 2013, classification of Marathi text documents was unexplored, so in work [11] various classification methods are compared for Marathi Text. After testing Naive Bayes, Centroid Classifier, KNN, and modified KNN, results concluded that Naive Bayes is most efficient considering time and accuracy. Here, Marathi text documents were pre-processed using rule based stemmer and Marathi word dictionary without removing stop words [11].

After viewing Marathi, Hindi and Bengali language with perspective of Information Retrieval (IR) [12] propounded light and aggressive stemming approaches. By applying some aggressive stemmers enhanced recovery effectiveness can be obtained. A significant performance contrasts were found after comparison betwixt no stemming and stemming indexing schemes. When an aggressive stemmer is applied, related improvements determined were approximately 18% for Bengali language, around 28% for Hindi language and approximately 42% for Marathi language as compared to a no stemming approach. In evaluation of these stemming technologies, FIRE 2008 Test collection and two language independent indexing methods i.e. n-gram and trunc-n are used. To expedite IR operation, two algorithmic stemmers were exhorted. First one is to remove inflectional suffixes and another is to remove frequently occurring derivational suffixes [12].
Table 2.1. Various text summarizers for Indian language

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Method</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bengali Language</td>
<td></td>
</tr>
</tbody>
</table>
| 1     | Text Summarization using Extractive Method. | 1) It scored the files of content in which query terms are having the highest frequency.  
2) The summary of text documents produced on the basis of query terms by applying vector-space-term-weighting. [5] |
| 2     | Text Summarization using Extractive Method. | 1) It decided the information on sentiments in the input text.  
2) K-mean is used for clustering of theme.  
3) It applies graph representation at relational level.  
4) Aggregation is accomplished by clustering of theme and by applying the graph representation.  
5) This approach used page ranking for selection of appropriate sentences [6]. |
| 3     | Text Summarization using Extractive Method. | 1)Following two important features are used to rank sentences:  
   - Thematic term  
   - Position.  
2) It uses TF-IDF techniques [7]. |
|       | Kannada Language |         |
| 1     | Text Summarization using Extractive Method. | 1)It is based on Information Retrieval (IR)  
2) This approach processes the input text and determines which lines are appropriate and which lines are not appropriate.  
3) This system uses command based interaction .[8] |
| 2     | Text Summarization using Extractive Method. | 1) It takes classified Kannada documents from online web sources.  
2) Thematic words are identified.  
3)This approach uses Inverse-Document-Frequency(IDF) method with Term-Frequency (TF) [9] |
| 3     | Text Summarization using Extractive Method. | 1) This approach is based on statistical.  
2) This method make easy to identify the exact topic. [10] |
<p>| 4     | Text Summarization using Extractive Method. | 1) It uses GSS coefficient and TF-IDF. |
|       | Punjabi Language |         |
| 1     | Text Summarization using | 1) Each line of input text is treated as vector. |</p>
<table>
<thead>
<tr>
<th>Extractive Method.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2) It uses GSS coefficient and TF-IDF [13].</td>
</tr>
<tr>
<td>3) Weight-age of sentence is determined by applying regression (weight learning method). [14] [15] [16]</td>
</tr>
</tbody>
</table>

### Hindi Language

<table>
<thead>
<tr>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text summarization based on statistical method</td>
</tr>
<tr>
<td>1) This approach is based on statistical.</td>
</tr>
<tr>
<td>2) It uses Language based approach along with heuristic approach.</td>
</tr>
<tr>
<td>3) It gives weightage and ranking to lines. [17]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Summarization using Extractive Method</td>
</tr>
<tr>
<td>1) This approach uses genetic algorithm and neural network for extracting the summary.</td>
</tr>
<tr>
<td>2) Selection of sentences is depending on the score.</td>
</tr>
<tr>
<td>3) Each sentence score is calculated on the basis of feature extraction.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text summarization is using fuzzy-logic</td>
</tr>
<tr>
<td>1) Fuzzy logic is used to extract important sentences.</td>
</tr>
<tr>
<td>2) Summary is generated by extracting the highest score sentences.</td>
</tr>
<tr>
<td>3) 20% compression ratio is used to generate document summary</td>
</tr>
<tr>
<td>4) After extracting the summary sentences are organized in the original order.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text summarization is based on rule based approach.</td>
</tr>
<tr>
<td>1) Handcrafted rules are developed to generate the summary.</td>
</tr>
<tr>
<td>2) System generates the extractive summary.</td>
</tr>
<tr>
<td>3) It does not include the semantic analysis of the Hindi text.</td>
</tr>
<tr>
<td>4) Summary of the Hindi text obtained from only single document.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) This approach uses statistical &amp; linguistic feature to extract the sentences.</td>
</tr>
<tr>
<td>2) Genetic Algorithm is also used [22].</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Summarization using Abstractive Method.</td>
</tr>
<tr>
<td>1) Rich Semantic Graph technique is used to generate summary [23].</td>
</tr>
</tbody>
</table>

### Tamil Language

<table>
<thead>
<tr>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Summarization using Extractive Method.</td>
</tr>
<tr>
<td>1) Semantic graph technique is used.</td>
</tr>
<tr>
<td>2) It recognizes Subject.</td>
</tr>
<tr>
<td>3) Object predicate from individual lines for making semantic-graph of document [13].</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Summarization using Extractive Method.</td>
</tr>
<tr>
<td>1) Syntax of language neutral applied to condense the text documents [13].</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Summarization using</td>
</tr>
</tbody>
</table>
Extractive Method. summarization.  
2) It used statistics of frequency of words.  
3) To score the sentences, weight-age of positional term are calculated by string pattern [18].

Malayalam Language

1. Text Summarization using Extractive Method. 1) It uses Machine Learning Approaches to generate the summary (Like Naive Bayes, Neural Network & Hidden Markov Model (HMM)) [19].

2. Text Summarization using Extractive Method 1) Maximum Marginal Relevance (MMR) techniques used with successive Threshold [20].

3 Text Summarization using Extractive Method. 1) It uses TF-IDF techniques [21].

Table 2.1 shows the existing summarizers method and approach used for Indian languages.

3 Method-Wise Techniques Used In Automatic Text Summarization

The short explanations about each approach are discussed in following section. Fig.3.1 shows various approaches used for summarization.

1 Abstractive text summarization:

Abstractive text summarization creates summary by understanding main concepts from source document thus expressing it in a clear natural language. Abstractive techniques are grouped into two classes [24].

Structured Based Approach:

This approach includes most relevant information from document by using intellectual schemes such as pattern, extraction rules. Structured based approach uses other structures such as ontology, tree, lead and body phrase structure to extract relevant information [25].
Semantic Based approach:

In this approach [24], documents are represented semantically and are go through Natural Language Generation (NLG) system. This method identifies verb phrases, noun phrases by examining linguistic data.

2 Extractive text summarization:

Extractive summarization extracts significant sentences from document by maintaining minimum redundancy of information in summary. In these methods, the most important regions of a document such as sentences, paragraphs and headline are given more weight age.

Fig. 3.1 Classification of various approaches used for summarization

3.1 Abstractive Summarization Methods:

3.1.1 Structured Based Methods

3.1.1.1 Tree Based:

The method uses language generation for producing brief summary. By using [40] shallow parser similar sentences are pre-processed using and then sentences are mapped to predicate-argument structure. With help of the content planner and theme intersection algorithm determine common phrases by comparing the predicate-argument structures. It uses units of the given document read and easy to summarize
3.1.1.2 Template Based:

In this approach, a topic is represented as a set of related approaches and implemented as a template containing slots and fillers. The templates are filled with important text snippets extracted by the Information Extraction systems. These text snippets are used to generate meaningful, informative multi document summaries by using IE based multi document summarization algorithm. This approach works just if the summary sentences are as of now present in the source documents. It can't deal with the functions if a multi document requires data about similarities and contrasts over various documents [26].

3.1.1.3 Ontology Based Method:

Ontology method is abstractive based summarization approach. Ontology in summarization is to increases interpreted concepts and joint relations between the concepts. Identifying the related concepts and finding out undetermined relations is important tasks in ontology-based summarization. Ontology approach present and determine the vocabulary used. It uses dictionary which is used in language. The result of ontology is a terminology agreement between users community. This agreement specifies which term is used to represent a concept in order to avoid ambiguity. This is nothing but vocabulary normalization. When a concept gives two synonym terms, the normalization process selects one of those labels which are from the concept. Finally summarization process generates a summary with preserve important information and shorter than original document [27].

3.1.1.4 Lead and Body Phrase Method:

This approach is abstractive based summarization. This method finds out lead sentences from the document. Whichever same chunks are there, are searched in body and lead phrases. By using similarity metric, repeated phrases are identified. Then substitution of body phrases takes place when it appears in body [40].

3.1.1.5 Rule Based Method:

The Rule Based methodology is based on an abstraction scheme. It generates summaries from clusters of news articles on same story. This method represents the summary in term of categories and list of forms. It uses Content selection module which selects the best candidate from the generated information extraction rules to answer one or more forms of a category. At the end generated patterns are used for generating summary sentences [40].

3.1.2 Semantic Based Methods:

3.1.2.1 Multimodal Semantic Model:

Multimodal document contains both text and images. Knowledge illustration concept
utilized in this model which is based on objects is constructed initially. With the help of information density metric contents are rated [40]. Parser stores expression which expresses the concepts and relationship. Concepts and relationships searched using metric which is based on applicability of concepts, number of relationships with different concepts and the number of expressions showing the existence of concept in the present document.

3.1.2.2 Information Item Based Method:

This approach provides abstract illustration of source documents, instead of from sentences of source documents. This framework based on three modules: Information Item retrieval, sentence generation, sentence selection and summary generation. First with parser Information Item retrieval module gives syntactic analysis of text and the triple which is verb’s subject and object. In sentence generation module, a sentence is directly generated from first module using a language generator, the NLG [40]. Sentence selection module ranks the sentences based on their average. Document Frequency (DF) score. Finally, a summary generation step take place which include dates and locations for the highly ranked generated sentences.

3.1.2.3 Semantic Graph Based Method:

This abstractive based approach contains three phases as the first Phase gives the given document semantically using Rich Semantic Graph (RSG). RSG used to create a semantic graph. In RSG, graph nodes nothing but [39] verbs and nouns of the input document with edges corresponding to semantic and topological relations between them. The second phase reduces the source document to more reduced graph using some Analytical rules. And the third Phase generates the abstractive summary from the reduced rich semantic graph. This stage accepts a semantic representation as RSG and creates the summarized text [23].

3.2 Extractive Based Methods:

3.2.1 Term Frequency-Inverse Document Frequency Method:

This is statistical method. This method is used to judge the importance of a term in a document. Term frequency is used to find the number of times a term occurs in a document. The term frequency of words like “the” can be very high. Inverse document frequency is calculated as the log of total number of documents divided by total number of documents in which the term occurs [29]. Inverse document frequency of a term can be low even if its term frequency is very high.

3.2.2 Cluster Based Method:

A document can address many documents in itself. They are normally broken up explicitly or implicitly into sections. Just like the document, summaries should
address different “themes” appearing in the document. If the document collection for which summary is being produced is of totally different topics, document clustering becomes very important to generate a meaningful summary.

Already clustered documents are input to the summarizer. Each cluster is treated as a theme. After clustering, sentences are selected from each cluster to give the summary. This is done so that all the themes of a document are covered in the summary. First factor for sentence selection is to determine how similar the sentences are to the theme of the cluster. Second factor is to locate the sentence in the document. The last factor is to calculate how much similar the first sentence is to document [30]. Overall score of sentence is given by the equation:

\[
S_i = W_1 \times C_i + W_2 \times F_i + W_3 \times L_i \tag{1}
\]

Where

\(S_i\) – score of sentence \(C_i\).

\(F_i\) - scores of the sentence \(i\) based on how much similar the sentences are to the theme of the cluster and first sentence of document respectively.

\(L_i\) - score of the sentence based on its location in the document.

\(W_1, W_2\) and \(W_3\) – weights

### 3.2.3 Graph Theoretic Approach:

In this technique, a graph is constructed. The nodes of the graph represent the sentences. The edges of the graph symbolize connection between the sentences which share the same words. The nodes which have more edges contain important sentences and are given more priority for summarization [31].

### 3.2.4 Machine Learning Approach:

In this method, the summarization task can be seen as a two-class classification problem. Sentences are grouped into summary sentences and non-summary sentences. The summarizer is trainable, the training data-set and their extractive summaries are used for reference. [32][33]
3.2.5 Latent Symantec Analysis (LSA) Method:

Singular Value Decomposition (SVD) is one of the most powerful tools to find principal orthogonal matrix in multidimensional information. [34] This technique is called LSA since SVD can be applied to document word matrices, collect documents that are semantically identified with each other however they do not share common words between them. Cosine similarity between rows (vectors) is calculated. Those vectors which have cosine similarity near about 1 have the most similar words, and those vectors whose cosine similarity is near about 0 have most dissimilar words. This technique extracts topic related sentences and word with essential contents from documents.

3.2.6 Text Summarization with Neural Networks:

In this method, each document is converted into a list of sentences. Each sentence is represented as a vector \([f_1, f_2... f_7]\), composed of 7 features. The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. Once the network has learned the features that must exist in the summary sentences, we need to determine the trends and relationships among the features that are intrinsic in the majority of sentences. This can be accomplished by the feature fusion phase, which consists of two steps: 1) eliminating uncommon features; and 2) collapsing the effects of common features [35].

3.2.7 Automatic Text Summarization Based On Fuzzy Logic:

This method considers each characteristic of a text such as sentence length, similarity to title, similarity to key word and etc. as the input of fuzzy system. All the rules required for summarization are also the input into the knowledge base. Each sentence gets a score ranging from zero to one as they are fed. This value determines the importance of sentences for summary generation. The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High h) and very high (VH). The important sentences are extracted using if-then rules according to the feature criteria [36].
3.2.8 Query Based Extractive Text Summarization:

In this approach, the summary is produced based on the query. Sentence scores are calculated based on the frequency of terms. Sentences which contain query phrases are scored higher than the sentences which contain single words from the query, and these are scored more than the sentences which do not contain any query term. Sentences with high scores are part of output summary [37].

4 Proposed Work

The proposed summarization system is categorized into three segments: pre-processing, sentence scoring and summarization.

Input to a summarization process can be one or more text documents. In single document summarization only one document is given as input, it is called single document text summarization but in multi-document summarization a group of related text documents is the input. Fig. 4.1 shows the flow of proposed text summarization system.

![Flow of the proposed text summarization system](image)
1 Pre-processing:

The preprocessing step consists of tokenization, stop-word removal, and stemming.

Tokenization: Every word is considered as a token. Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Each sentence is broken down into tokens and words of different language or punctuation might be removed.

Stop words removal: The words which occur frequently but do not add any semantic value to the document are termed as stop-words. In Marathi words आणि (And), तरी (Or), परिंतु (But), म्हाून (As) etc. are used frequently in sentences which have little significance in the implication of a document. These words can simply be removed for classification process.

Stemming: A single word can have many different forms. These words have to be converted to their original form for e.g. “stemmed”, “stemmer”, “stemming” are all from the root word “stem”. Stemming can be helpful for calculating the word frequency of a root word.

2 Sentence Ranking and Summarization

After pre-processing, the sentences are ranked based on four important features: Frequency, Position value, Cue words and Skeleton of the document.

Frequency: Frequency is the number of times a word occurs in a document. If a word’s frequency in a document is high, then the word is important for the document.

Positional Value: The positional value of a sentence is calculated by assigning the highest value to the first sentence and the lowest value to the last sentence of the document. The sentence with higher value is more important.

Cue Words: Words such as therefore, hence, lastly, finally, meanwhile or on the other hand are cue words. These words are connective expressions. The sentences
containing cue words are usually very significant.

Skeleton of the Document: The words in title and headers in document are considered as skeleton of the document. Those words are having some additional weights in sentence scoring for summarization.

Sentence Scoring: The final score of sentence is calculated by a Linear Combination of all above features (i.e. frequency, positional value, weights of Cue Words and Skeleton of the document).

3 Summary Making

In our proposed work, sentences are ranked on the basis of their scores and selecting K-top ranked sentences. In this value of K is set by user. Summary is produced after ranking these sentences. The sentences in the summary are reordered based on their presence in the original text to increase the readability of the summary; for example, the sentence which take place first in the original text will appear first in the summary.

5 Conclusion

Natural Language Processing (NLP) is the area where Automatic Text Summarization is a well-known task. It can broadly be divided into two classes: Abstractive and Extractive summarization. Abstractive summary deals with understanding the semantic relationship between the texts whereas; extractive summary deals with directly extracting important sentences from the document and then creating the summary. Therefore, abstractive summarization is little more tedious to carry out than extractive summarization. In the proposed approach feature extraction is primary phase, after those sentences are scored. In the last phase higher ranked sentences are selected as a summary. In this paper we have briefly illustrated automatic text summarization techniques for several Indian regional languages (Hindi, Tamil, Bangla, Kannada, English and Panjabi etc). However there are various techniques that have been proposed for several languages but a very few effort have been taken for Marathi text.
Reference


14. Gupta, V., Lehal, G.S.: Automatic Text Summarization for Punjabi Language,


40. Khan, A., Salim.: A Review On Abstractive Summarization Methods - Faculty of Computing, Universitiy Teknologys Malaysia, 81310, Skudai, Johor, Malaysia