International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I6P107

Volume 69 Issue 6, 44-49, June 2021

Original Article

Text Summarization using Extractive Techniques for Indian Language

Manasi Chouk[1], Neelam Phadnis[2]

[1]*PG Student, Department of Computer Engineering, Shri. L.R.Tiwari College of Engineering, Mumbai, Maharashtra, India.*
[2]*Assistant Professor, Department of Computer Engineering, Shri. L.R.Tiwari College of Engineering, Mumbai Maharashtra, India.*

Received Date: 05 May 2021
Revised Date: 10 June 2021
Accepted Date: 16 June 2021

Abstract *- Over the past decade there have been significant advances in Natural Language Processing and Machine Learning technologies. Research work is currently being done to summarize large text documents which are difficult to summarize manually. In Text summarization process the most important text has been selected and forms summaries. The purpose of the summarization system is to generate a short abstract and allows readers to understand the objective of text instead of reading original document.*

This paper describes various extractive and abstractive techniques can be used to summarize large Marathi text. Summarization techniques be categorized as: Abstractive summarization and Extractive summarization techniques. Various text summarization systems are implemented for English and other languages. However very few techniques are available for Indian language such as Marathi.

Keywords *- Extractive technique, Marathi Language, NLP, Text Summarization, TF-IDF, Text Rank.*

I. INTRODUCTION

Natural Language Processing (NLP) is a group, which has computational method to analyse and represent text occurring in the document. The linguistic analysis is done at various levels for a huge span of tasks or applications like text summarization.

The aim of text summarizer systems is to generate the short abstract with semantics from the given text. The intention is to create a clear and easy to understand summary with main points from the document. The advantage is that it reduces the reading time.
Main Approaches for Automatic Text Summarization are

A. Extraction-Based Summarization

An extractive technique comprises of choosing main words, sentences, etc. from the original document and combining them into short abstract form. It involves extracting important words or sentences from the input text and merging them to form a short abstract text. The extracting is done in accordance with defined metrics and no change will be made to the input text. These summarizers are mostly works on sentence score from the input text. The technique used will be either statistical or linguistic approach.

B. Abstraction-Based Summarization

An Abstractive summarization is a technique of getting the important text from the original document and expressing these text in clear natural language. These abstractive techniques creates new words and sentences that are relevant to the input text.

II. LITERATURE REVIEW

This section describes past research work done in the field of Text Summarization for Indian Languages such as Hindi, Marathi, Gujarati, Bengali etc. Extensive work has been done in English Language as it is easy to implement because of availability of pre-defined libraries which makes pre-processing of text easier. However for regional language, pre-processing is a difficult task. Hence, all the steps has to be implemented separately. Currently most of the work has been done using extractive techniques for regional languages compared to abstractive technique.

Oguzhan Tas and Farzad Kiyani has describes single and multi-document extractive text summarization techniques for English language. He discussed Query based and generic summarization method which includes Bayesian Classifier, Hidden Markov Model, Neural Networks and Fuzzy Logic etc. [1]

Later Sheetal Shimpikar and Sharvari Govilkar discussed various text summarization techniques for Indian regional languages and have compared them. Also discussed different types of extractive and abstractive text summarization technique. [2]

Vaishali V. Sarwadnya and Sheetal S. Sonawane developed automatic text summarizer using extractive technique for Marathi language. For this they have used graph based model. [3]

Akash Ajampura Natesh and et al. proposed a summarization system using graph based approach. In this method, author used a comcept in which they have computed important words from the sentences and checks the relation between the words. Futher the metric weighs the important sentences from the input document.[4]

Madhurima Dutta et al. proposed a method, in which a representative shorter version of text has been created from the whole text. It is a graphical representation in which important sentences were found out by a technique called infomap clustering. [5]

Rahim Khan et al. proposed extractive based summarization using K-Means Clustering with TF-IDF (Term Frequency-Inverse Document Frequency) for summarization. This paper describes the idea of true K and using that value of K divides the sentences of the input document to present the final summary. Furthermore, they have combined the K-means, TF-IDF with the issue of K value and predict the resulting system summary which shows comparatively best results. [6]

Virat V. Giri et al. proposed a single document multi news Marathi extractive summarizer. This system is used to summarize the single Marathi document with multi news by retaining the relevant sentences based on statistical and linguistic text features. [7]

Shubham Bhosale et al. proposed an algorithm to extract keyword automatically for summarization from e-newspaper articles. The keyword extraction algorithm is used to find the top scored words and by using this data the summarization module produces summarized article. [8]

Apurva D. Dhawale et al. proposed a system which pre-process the Marathi text. In step one , the input is extracted, then the length of text is calculated, then tokenized, sentence length is calculated, special symbols are removed, then the frequency count of the word is taken as a statistical value and key value pairs are formed for further processing. [9]

Anushka Chaudhari et al. proposed extractive text summarization system using neural network. for this they have used Recurrent Neural Network (RNN) which can be used to perform calculations on sequential data. Also the translation of the Marathi text to English is done using the Google translate API. [10]

Jayshree Arjun Patil et.al describes a brief introduction to NER and discussed various approach and challenges faced for Indian languages. [11]

Nita Patil et.al discussed various challenges and issues that can be faced during implementing NER system for Marathi language. [12]

Deepali K. Gaikwad et al. proposed a system which is used to generate Question and accepts Marathi text as input and processed the input by applying POS tagging NER and stemming and then generate the question as per the rules. The answers of the generated questions is the summary of the given input. [13]

Mudassar M. Majgaonker et al. included a system which evaluates a rule-based and an unsupervised Marathi stemmer. The rule-based stemmer uses a set of manually extracted suffix stripping rules whereas the unsupervised approach learns suffixes automatically from a set of words extracted from raw Marathi text. [14]

Sheetal Shimpikar et al. proposed abstractive text summarization technique for Marathi text using rich semantic graph method for Marathi language. The pre-processing phases included input validation, tokenization, stemming and morph analysis, POS tagging, and NER. Then the pre-processed sentences are given as input to rich semantic graph method. [15]

Manjula Subramaniam et al. presented a abstractive method for summarizing Hindi Text document by creating rich semantic graph(RSG) of original document and identifying substructures of graph that can extract meaningful sentences for generating a document summary. [16]

K. Vimal Kumar et al. , proposed a system which is focused on the Hindi language. The main idea of this summarization system was to identify the important sentences and to extract them based on its relevance with other sentences. [17]

Shohreh Rad Rahimi et al. discussed relationship between text mining and text summarization and also discussed some of the extractive approaches and their important parameters such as important sentences, identifying main stages in summarization process. [18]

ZHANG Pei-ying et al. proposed a sentences clustering based summarization approach. The proposed approach consists of three steps: first clusters the sentences based on the semantic distance among sentences in the document, and then on each cluster calculates the accumulative sentence similarity based on the multi features combination method, at last chooses the topic sentences by some extraction rules. [19]

Asha Rani Mishra et al. proposed a technique for extracting important information from the given input text using text modelling, key phrase extraction and summary generation. [20]

JIANG Xiao-Yu et al. proposed a technique to reduce the dimensionality of feature vector space and reduce the computing complexity of categorization, each document of the train set was summarized automatically and two approaches to text categorization based were proposed first is feature selection and categorization and second is weight feature. [21]

Taeho Jo et al. proposed text summarization algorithm using KNN (K Nearest Neighbor) where the similarity between feature vectors is computed considering the similarity among attributes or features as well as one among values. [22]

Eliseo Reategui et al. presents a mining tool which is used to extract graphs from texts, and proposes their use in helping students to write summaries. The tool is developed using a particular mining algorithm based on the n-simple distance graph model, in which nodes represent the main terms found in the text, and the edges represent adjacency information. [23]

Anish Jadhav et al. describes an algorithm which is based on both extractive and abstractive technique. They used neural network concept. Initially, important sentences are identified and merged together to form a document. The significance of a sentence is chosen with their semantic meaning of sentences. This shorter representation is then passed through an Encoder-Decoder model to generate a concise summary representing the whole article. The proposed model is capable of effectively creating a concise summary, which is semantically and linguistically correct. The proposed methodology focuses only on the relevant sentences and passes it to the Bi–Directional RNN for identifying and representing the core idea of the article. [24]

Ayush Agrawal et al. describes an algorithm that incorporates k-means clustering, term-frequency inverse-document-frequency and tokenization to perform extraction based text summarization. [25]

Nithin Raphal et al. explains the overview of the various processes in abstractive text summarization which includes data processing, word embedding, basic model architecture, training, and validation process. [26]

Huong Thanh Le et al. describes an approach for abstractive text summarization based on discourse rules, syntactic constraints, and word graph. Discourse rules and syntactic constraints are used in the process of generating sentences from keywords. Word graph is used in the sentence combination process to represent word relations in the text and to combine several sentences into one. [27]

Atif Khan et al. presented two main categories of abstractive text summarization i.e structured based approach and semantic based approach. [28]

Khushboo S. Thakkar et al. proposed unsupervised methods for automatic sentence extraction using graph-based ranking algorithms and shortest path algorithm. [29]

Chetana Badgujar et al. proposed a work which will mainly focused on graph based technique and analyse the details. [30]

Shashi Pal Singh et al. proposed Bilingual (Hindi and English) unsupervised automatic text summarization using deep learning. For this they have used restricted Boltzmann machine to generate a shorter version of original document without losing its important information and explores the features to improve the relevance of sentences in the dataset. [31]

III. EXTRACTIVE TEXT SUMMARIZATION

Extractive summarization means classifying main sentences from the document and generating group of sentences from the original documents;
The core of all extractive summarizers is based on three important steps:

- Implementing an intermediate representation of document.
- Sentence score based on representation
- Selecting the sentences for generating summary.

Two main types of representation methods are used: topic representation and indicator representation. Topic representation converts the content into an intermediate representation and interpret the topic explained in the text. In Indicator representation every sentence in a list indicates some features like length, position etc. in the document.

A. Topic Representation Approaches

a) Topic Words

This method identifies words that are related to the topic of the input text.

b) Frequency Driven approaches

This technique calculates frequencies of important words present in the document.

c) Latent Semantic Analysis

This technique extracts and represents semantic words from the input document.

d) Discourse Based Methods

In this methd, analysis of the semantic words is done and finding their relationship among words, which will generate summary.

e) Bayesian topic Models

It is a probabilistic method in which the information that is lost in other techniques is preserved and all topics are explained in detail.

B. Indicator Representation Approaches:

a) Graph Methods

In this approach, sentences are the vertices and edges indicates the similarity between two sentences using a connected graph. The documents are represented as a connected graph

b) Machine Learning

The various Machine learning approaches are applied for text summarization and classification techniques to obtain a summary.

IV. EXTRACTIVE SUMMARIZATION METHODS

A. TFIDF

Term frequency–inverse document frequency, is a statistical approach which reflects the importance a word in a set of document or corpus. In TFIDF method, the semantic information of word as well as uncommon words are preserved. It gives more importance to uncommon words rather than common words. TF of a word is defined as the term frequency of that word in that particular document and IDF of word in whole corpus of document. Value of tfidf is increased as the same word appear in the document with the offset by total number of documents in the data set which contains that same word. This is the most used term weighting technique.

B. Cluster Based Method

In Cluster based technique, the semantic nature of the input text is preserved and form a small set of sentences. These set is nothing but subjects, verbs, objects related to each sentence in a document. Using these sets, clusters are formed by considering similar information from the documents. The set of sentences is nothing but the sentences and used in summarization process. These sets are computed and a sequence of sentences related to the set forms asummary.

C. Graph Theoretic Approach

This is a graphical representation method in which sentence are nodes and edge is connection of two sentences. If two sentences has some common words, then their similarity is above threshold value. This representation has two results : first part partitioned containes different topics present in the documents and the second part is by the graph-theoretic method which is the important sentences are identified from the document. Text Rank is a text summarization technique which is used to generate Document Summaries. Text Rank uses an extractive approach and is an unsupervised graph-based text summarization technique. Text Rank Algorithm is inspired by Page Rank algorithm which is primarily used for ranking web pages in online search results.

D. Machine Learning Approach

In this approach ,for reference the training dataset is used and the summarization process is modelled as a classification problem: sentences are classified as summary sentences and the sentences which are not in summary are based on the features that they possess.

E. Using Neural Networks

In neural network, documents are converted sentences and these sentences are represented as a vectors. These vectors are composed of different features such as sentence length, number of thematic words in sentences, number of title words in sentences, sentence location etc.

F. With Fuzzy logic

It is an approach to variable processing that allows for multiple input values to be processed. It is designed to process by considering all possible information available and giving best possible result from the given input. In this technique,the characteristics of a text document are sentence length, similarity to title words present, sentence location in document are considered as a input to fuzzy system .Then, the next step is to enter all the rules required for summarization process in knowledge base system. Then, a range of values from zero to one is obtained for every sentence in the output. These values are based on sentence characteristics and the available rules in the knowledge base. The obtained value determines the degree of the importance of the sentence in the final summary.

V. ABSTRCTIVE SUMMARIZATION METHODS

In abstractive text summarization, multiple documents are taken as input and generates words or sentences which may be or may not be present in input document and creates summary. It records the main features of text document and generates summary. The output summary is looks like human generated summary. Very less work has been done in this area for Indian languages. The text summarization can be done using four processes: data processing, word embedding, model architecture and train-test evaluation model.

Following are the two types of Abstractive technique : Structured based and Semantic based .

A. Structured Based Method

Structured based approach primarily compute most important data from the input document. Important data includes features like rules to extract text, templates and structures like tree, ontology based, rule based.

a) Tree Based Method

In this method, a dependency tree is used to represent text of the document. This technique uses an algorithm for generating a summary. At the same time, the algorithms available for text selection changes accordingly from theme intersection. Various algorithms are used for content choice for parsed sentences. The outline is generated either with the help of a language generator or an associate degree algorithm.

b) Template Based Method

In this technique, entire document is represented using template. All linguistic patterns and extraction rules were matched to spot text snippets that are mapped into guide slots. These text snippets are the area unit which indicates the outline content.

c) Ontology Based Method

It is knowledge base method used to improve the summarization process. Ontologies based methods are convenient to use because it is restricted to same domain or topic. Every topic has its knowledge structure and it can be better represented by ontology technique. Ontology based methods are different in their specific approaches. Linguistic and NLP based approaches are used to reduce the sentences compressing and reformulation.

d) Rule Based Method

In this technique, classes are made from the input text.A rule-based information extraction module, content selection heuristics and one or more patterns are used for generating sentences. Extracting rules are used to identify same meaning verbs and nouns. Multiple candidate rules are selected and passed to summary generation module. In the last step, patterns are used for sentence generation. This technique generates the best summary but the disadvantage is the time required is more as rules and patterns are written manually.

B. Semantic Based Method

Semantic based approach make use of linguistic illustration of document which is then transfered to natural language generation system for identifying verb and noun phrases.

a) Multimodal Semantic Based Method

This method forms a relation among features which represents text and images in multimodal documents. Objects based knowledge representation model is implemented. In this model, features are the nodes and edges between these features shows the relation between them. Main features or concepts are ranked using information density metric. This metric is used to check the completeness of the features and these features are converted into sentences summary generation.

b) Information Item Based

In this technique , abstract text from the input document is used instead of sentences for creating summary. Information item is an abstract text of the logical information of the generated summary. The summary is generated from abstract information rather than the input document.

c) Semantic Graph Based

In this method, rich semantic graph is created and is used to generate the summary. This method comprises of three steps; in first step, input text is represented as rich semantic graph, the verb and noun are nodes and edges corresponds to semantic and topological relation between them. The second step is the reduction of the original graph into more concise graph by applying heuristic rules. In the third step, abstractive summary is generated from the reduced semantic graph. In this method, the summary generated is grammatically correct and less redundant.

VI. CONCLUSION

Text Summarization is an important field in NLP. Most of the research work has been done using extractive technique. For Marathi language, very few techniques are implemented because of the unavailability of resources. The combination of different types of features work differently for Marathi text. Hence , a single summarizer may not work for different type Marathi text. In future, our aim would be to add more features to extract the Marathi text. Also, we will try to use abstractive technique for the same.

REFERENCES

[1] F. K. Oguzhan Tas, A Survey Automatic Text Summarization, PressAcademia Procedia, 5(2017)(2017) 204-2013.

[2] S. Shimpikar and S. Govilkar, A Survey of Text Summarization Techniques for Indian Regional Languages, International Journal of Computer Applications (0975 – 8887), 165(11)(2017) 29-33.

[3] S. S. S. Vaishali V. Sarwadnya, Marathi Extractive Text Summarizer using Graph Based Model, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Vols. 978-1-5386-5257-2/18, 2018 IEEE.

[4] S. T. B. A. P. P. Akash Ajampura Natesh, Graph Based Approach for Automatic Text Summarization, IJARCCE, 5(2)(2016) 6-9.

[5] A. K. D. C. M. A. S. Madhurima Dutta, A Graph Based Approach on Extractive Summarization, ResearchGate, 2(2018) 1-9.

[6] Y. Q. S. N. Rahim Khan, Extractive based Text Summarization Using K-Means and TF-IDF, I.J. Information Engineering and Electronic Business, 3(2019) 33-44.

[7] D. M. a. D. K. Virat V. Giri, A Survey of Automatic Text Summarization System for Different Regional Language in India, ResearchGate, 6(2016) 52-57.

[8] M. D. J. M. V. B. P. A. D. Mr. Shubham Bhosale, Marathi e-Newspaper Text Summarization Using Automatic Keyword Extraction Technique, International Journal of Advance Engineering and Research Development, 5(3) (2018) 789-792.

[9] S. B. K. V. M. K. Apurva D. Dhawale, Automatic Pre-Processing of Marathi Text for Summarization, International Journal of Engineering and Advanced Technology (IJEAT), 10(1)(2020) 230-234.

[10] A. D. D. K. Anishka Chaudhari, Marathi text summarization using neural networks, International Journal of Advance Research and Development, 4(11) (2019) 1-3.

[11] M. P. B. G. Ms. Jayshri Arjun Patil, Review of Name Entity Recognition in Marathi Language, IJSART, 2(6) (2016) 497-499.

[12] A. S. P. A. B. V. P. NITA PATIL, Issues and Challenges in Marathi Named Entity Recognition, International Journal on Natural Language Computing, 5(1) (2016) 15-30.

[13] D. S. a. C. N. M. Deepali K. Gaikwad, Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer, IOSR Journal of Computer Engineering (IOSR-JCE),51-54.

[14] T. J. S. Mudassar M. Majgaonker, Discovering suffixes: A Case Study for Marathi Language, (IJCSE) International Journal on Computer Science and Engineering2716*, 2(8)(2010) 2716-2720.

[15] S. G. Sheetal Shimpikar, Abstractive Text Summarization using Rich Semantic Graph for Marathi Sentence, JASC: Journal of Applied Science and Computations, 5(12)(2018) 2381-2386.

[16] P. V. D. Manjula Subramaniam, Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method., International Research Journal of Engineering and Technology (IRJET), 02(02)(2015) 113-116.

[17] D. Y. a. A. S. K. Vimal Kumar, Graph Based Technique for Hindi Text Summarization, Researchgate, (2015)301-310.

[18] A. T. M. M. A. Shohreh RadRahimi, An Overview on Extractive Text Summarization, IEEE, (2017) 54-62.

[19] L. C.-h. ZHANG Pei-ying, Automatic text summarization based on sentences clustering and extraction, IEEE, 2009.

[20] V. P. P. K. Asha Rani Mishra, Extractive Text Summarization – An Effective approch to Extract information from text, IEEE, (2019) 252-255.

[21] F. X.-Z. W. Z.-F. K.-L. JIANG Xiao-Yu, Improving the Performance of Text Categorization using Automatic Summarization, IEEE, (2009) 347-351.

[22] T. Jo, K Nearest Neighbor for Text Summarization using Feature Similarity, IEEE, 2017.

[23] M. K. M. D. F. Eliseo Reategui, Using a Text Mining Tool to Support Text Summarization, IEEE, (2012) 607-609.

[24] R. J. S. F. M. S. S. Anish Jadhav, Text Summarization using Neural Networks, IEEE, (2020).

[25] U. G. Ayush Agrawal, Extraction based approach for text summarization using k-means clustering, International Journal of Scientific and Research Publications, 4(11) (2014).

[26] H. D. ,. P. D. Nithin Raphal, Survey on Abstractive Text Summarization, International Conference on Communication and Signal Processing, Vols. 978-1-5386-3521-6, no. April 3-5, 513-517, (2018) IEEE.

[27] T. M. L. Huong Thanh Le, An approach to Abstractive Text Summarization, 2013 International Conference of Soft Computing and Pattern Recognition (SoCPaR), Vols. 978-1-4799-3400-3/13, (2013) 371-376, IEEE.

[28] N. S. ATIF KHAN, A Review on Abstractive Summarization Methods, Journal of Theoretical and Applied Information Technology, 59(2014) 64-72.

[29] D. R. V. D. M. B. C. Khushboo S. Thakkar, Graph-Based Algorithms for Text Summarization, IEEE, (2010) 516-519.

[30] V. J. T. G. Chetana Badgujar, Abstractive Summarization using Graph Based Methods, IEEE, (2018) 803-807.

[31] A. K. ,. A. M. ,. S. S. Shashi Pal Singh, Bilingual Automatic Text Summarization Using Unsupervised Deep Learning, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) , Vols. 978-1-4673-9939-5/16, 1195-1200, 2016.