Automatic Unsupervised Extractive Summarization of Marathi Text Using Natural Language Processing

Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna
Dept. of Computer Science & IT, Dr. B. A. M. University, Aurangabad, India.

Abstract: *The manual document summarization of any document is very tedious job and such summarized documents may differ and lead to diverse results. To avoid this problem, automatic text summarization plays very vital role. An Automatic Text Summarization (ATS) System gives the summary of any document without changing its meaning and with less sentences. There are a lot of ATS systems available for various languages. In India, variety of regional languages are spoken and people do prefer to read the books, news articles, journals, magazines in their regional language only. Here emanates the summarization of regional language of Maharashtra i.e. Marathi. The Marathi text summarization is now imperative task and it is equally significant for the young one's too. The System is developed according to the need of students, who are willing to give competitive exams and they do need the awareness of current affairs in terms of daily news articles. This system summarizes Marathi news articles using machine learning algorithm.*

Keywords: *ATS (Automatic Text Summarization), Gensim, Unsupervised Extractive Text Summarization*

Date of Submission: 26-12-2020 Date of Acceptance: 08-01-2021

I. Introduction

The Natural Language Processing is a demanding and core arena now a days. There are many applications of NLP with very vital roles. One of them is text processing of various novels, news, books, magazines etc. The people are now more aware of using e-news papers and that too in their regional language to minimize the reading time, carrying hard copy of news paper. The best way to use e-news paper may be summarized news to get exact information in minimum words without changing its meaning. There are two techniques which may be followed for text summarization. One is Abstractive and other is Extractive Text Summarization. We are demonstrating Unsupervised Extractive Text Summarization which follows TextRank Algorithm.

II. Literature Review

There is a plenty of work done in NLP using multiple foreign and Indian languages. Also, many researchers proved that extractive summarization works very well as it doesn't change the sentence construction and POS methods. The TextRank algorithm is used for numerous languages but while using python, Gensim library works efficiently with Marathi Text. We have developed a syatem which summarizaes marathi e-news articles for the students who are preparing for the competitive examiations.

Compressing the text without changing its meaning is called as text summarization. There is lot of difficulty in getting a higher efficiency in text summarization of Indian regional languages. A lot of work has been found for English language which has better results. So, it is needed now to focus on regional languages which are used in various fields for various purposes.[1]

Machine learning approaches use machine learning algorithms for the generation of summaries. The summarization process is dealt with as a classification problem, based on the features, there are two types of classification: summary and non-summary.[2]

We are focussing on educational, Political and sports news for summarization, which will be helpful for students who are appearing for competitive exams. This paper explores the pre-processing techniques for Marathi e-news articles.

Pooja Bolaj,SharvariGovilkar[5] developed a text classification system for Marathi documents using supervised learning methods & ontology based classification technique which classifies Marathi documents belonging to Festival class i.e. Diwali.

Deepali K. Gaikwad, Deepali Sawane and C. Namrata Mahender, seveloped a system for rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer. The paper shows technique which is used for generation of the appropriate question on given input/text.[6]

Yogeshwari V. Rathod [7] used sentence ranking algorithm to generate summary of Marathi news articles by extractive method. It gives effective summary in less time and with least redundancy.

DOI: 10.9790/0661-2206022125

Shraddha A. Narhari, RajashreeShedge [8] proposed a text categorization of Marathi documents using LINGO & PCA algorithm. They proved this with improved results.

Jaydeep Jalindar Patil, Prof. NagarajuBogiri[9] used LINGO [Label Induction Grouping] algorithmfor improving results efficiently inmarathi text documents.

Prakhar Sethi, Sameer Sonawane, SaumitraKhanwalker, R. B. Keskar [10] developed a system to Overcome the limitations of the lexical chain approach to generate a good summaryusing the WordNet thesaurus, pronoun resolution for news articles.

N. Dangre, A. Bodke, A. Date, S. Rungta, S.S. Pathak [11] proposed a System for Marathi News Clustering using Cluster algorithm to collect relevant Marathi news from multiple sources on web which results in enabling rich exploration of Marathi contents on web.

Mr. Shubham Bhosale, Ms. Diksha Joshi, Ms. VrushaliBhise, Prof.Rushali A. Deshmukh [12] proposed a system for Marathi newspaper text summarization using Ranking algorithm which gives average of 30% to 40 % size of original article.

Anishka Chaudhari, Akash Dole, Deepali Kadam, proposed a system which translates Marathi dataset to English using Google Translate API and then summarizes news articles using a bi-directional encoder-decoder LSTM model. The resultant summary is again translated to Marathi using Google Translate API.[13]

Features can be developed based on the perspective levels and divided into five categories such as word, sentence, summary, readability, and source-side features. Evaluation metrics are used to compute the word segmentation by using F score and ROUGE versions [14].

Sentence ranking formula is used to maintain the combination between summarization and information [15]. Word's information can be calculated by using logarithmic equations.

The research work [16–18] takes input sentence from the document, and preprocesses it by separating words, removing stop words, and then the last step is stemming.

This proposed system [19] elaborates the most important steps in this system approach are feature extraction, scoring and graph generation. This system can be used in various fields like education, in search engines to improve their performances, for Marathi news clustering, Question generation purpose and many other application oriented areas, etc.

The method in this paper [20] proves that the precision, recall and f-measure values of Decision Module using five features are better as compared to Decision Module using nine features - the only exception being the Type 2 documents. The achieved results of Decision Module are a promising start toward further studies. The summarizer can have a wide range of applications – Summarizing reports, articles, selecting most relevant document out of a large number of related documents etc.

The system [21] proves that Cohesion feature is grammatical and lexical linking within sentences that hold the sentence and provide meaningful text to end user without changing source text idea. Cohesion Feature through generated summery compared with commercial online summarizer or Microsoft summarizer tool therefore by adding cohesion feature, The text overloading problem is solved and the effectiveness of summery was increased.

III. Proposed System

In this research, we are using extractive based approach using Text ranking algorithm where the document is read first, its length is calculated, and it would generate a summary which gives us important sentences according to the requirement of the user. The relevant literature shows that there are many methods & algorithms suitable for Text processing and text summarization as the digital text is gaining importance day by day. The result may vary depending on the language chosen and the selected algorithm. Marathi is considered as an Indo-Aryan language. [3]

The experiments are done on the unstructured data to convert it into more comprehensible form that is structured data. The text rank algorithm follows the steps shown in fig.1 to get summary.[4]

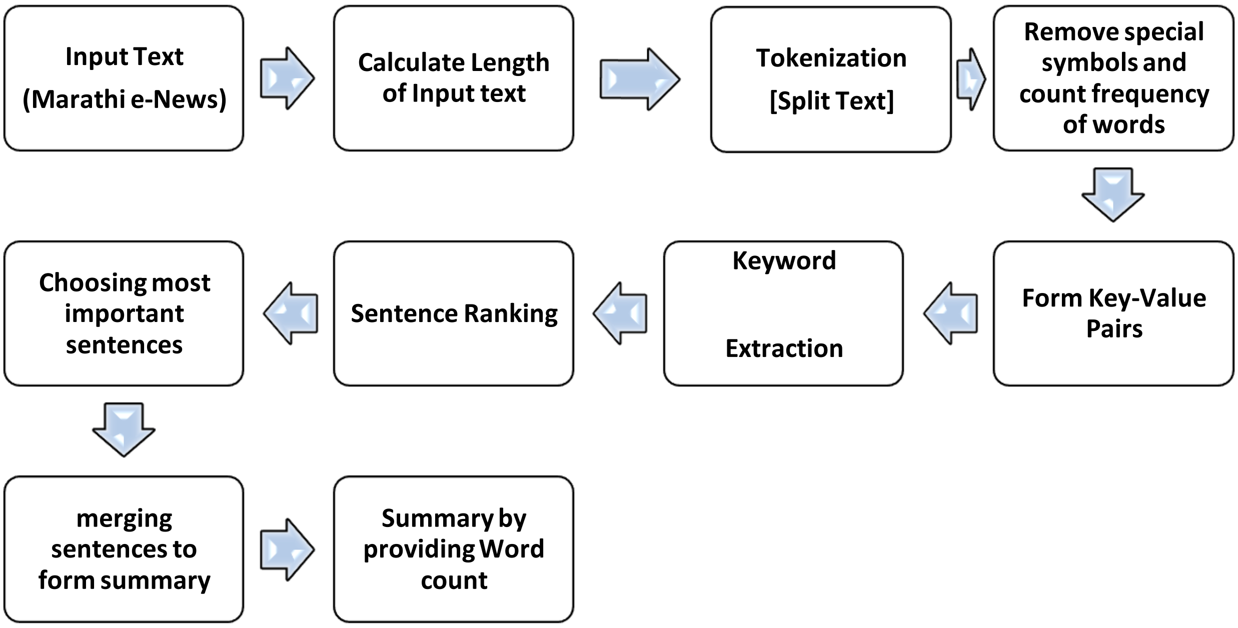

Input Text
(Marathi e-News)
Calculate Length of Input text
Tokenization
[Split Text]
Remove special symbols and count frequency of words
Form Key-Value Pairs
Keyword
Extraction
Sentence Ranking
Choosing most important sentences
merging sentences to form summary
Summary by providing Word count

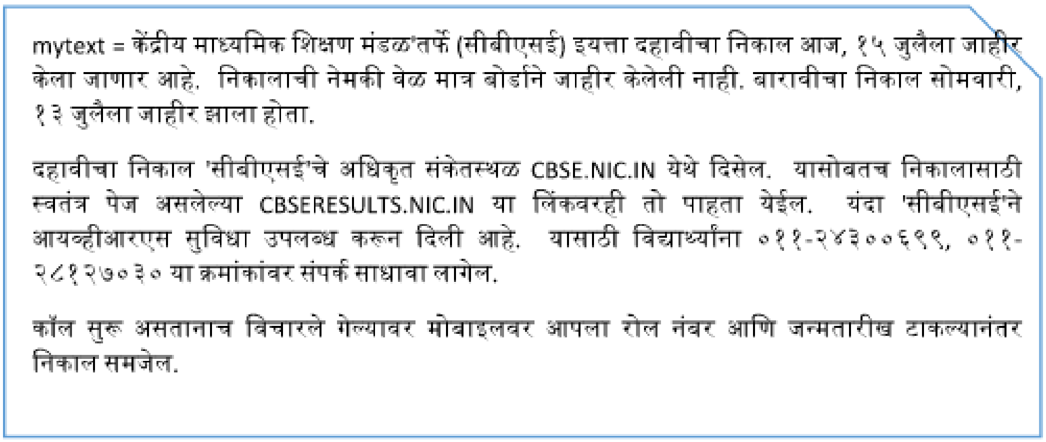
Fig1: Proposed system for Marathi e-news summarization

The proposed system shows the process of generating summaries of Marathi text. We are using dataset from github which has 1135 Marathi e-news articles. These are txt files and the contents are then passed on as an input for summarization. The figure above represents the proposed system for Marathi e-news summarization.

IV. Experiment, Evaluation and Result

This approach summarizes input text by using extractive summarization method. Here, there is a need of summarizing text by counting number of words, or by providing ratio which will be the value between 0 to 1. If user uses the first method i.e. give count of words you require in summary, then the summary will be of the given count only.

In second method, if user is providing 0.2 value, it will give 20% summary of total input text. Likewise, if 0.9 value gives 90% of text in output. Ratio also plays important role in Gensim library.

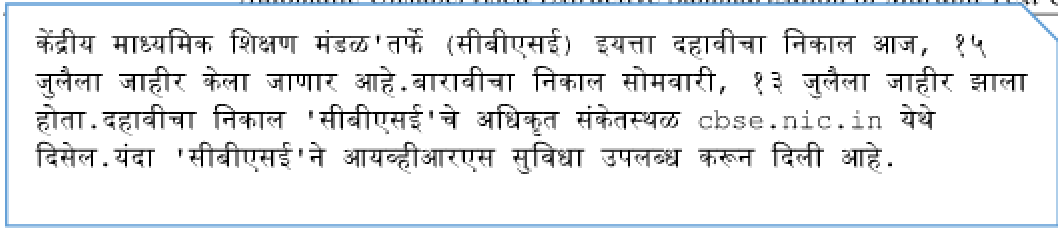


mytext = केंद्रीय माध्यमिक शिक्षण मंडळ'तर्फे (सीबीएसई) इयत्ता दहावीचा निकाल आज, १५ जुलैला जाहीर केला जाणार आहे. निकालाची नेमकी वेळ मात्र बोर्डाने जाहीर केलेली नाही. बारावीचा निकाल सोमवारी, १३ जुलैला जाहीर झाला होता.

दहावीचा निकाल 'सीबीएसई'चे अधिकृत संकेतस्थळ CBSE.NIC.IN येथे दिसेल. यासोबतच निकालासाठी स्वतंत्र पेज असलेल्या CBSERESULTS.NIC.IN या लिंकवरही तो पाहता येईल. यंदा 'सीबीएसई'ने आयव्हीआरएस सुविधा उपलब्ध करून दिली आहे. यासाठी विद्यार्थ्यांना ०११-२४३००६९९, ०११-२८१२७०३० या क्रमांकांवर संपर्क साधावा लागेल.

कॉल सुरू असतानाच विचारले गेल्यावर मोबाइलवर आपला रोल नंबर आणि जन्मतारीख टाकल्यानंतर निकाल समजेल.

Input Text provided:
The Extractive text summarization by providing count of words is as follows:



केंद्रीय माध्यमिक शिक्षण मंडळ'तर्फे (सीबीएसई) इयत्ता दहावीचा निकाल आज, १५ जुलैला जाहीर केला जाणार आहे.बारावीचा निकाल सोमवारी, १३ जुलैला जाहीर झाला होता.दहावीचा निकाल 'सीबीएसई'चे अधिकृत संकेतस्थळ cbse.nic.in येथे दिसेल.यंदा 'सीबीएसई'ने आयव्हीआरएस सुविधा उपलब्ध करून दिली आहे.

Word count = 40
The following table shows the comparison between lengths of input text and output text by using word count:

Table1: Comparison of length of input text and summarized text using word count.

Length of Input text	**Length of Output text by using word count**
607	278

Graphical representation of the table1 is shown below:

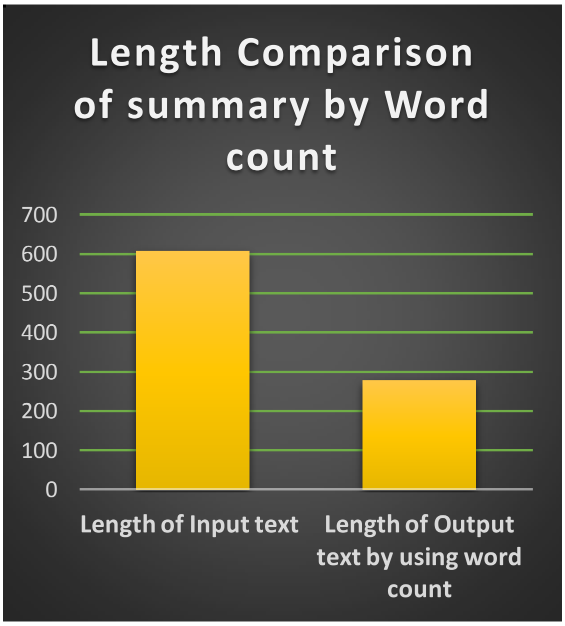


Fig2: Comparison of length of input text and summarized text using word count.

As shown in the above figure, the input and output length are calculated and compared to show the summary by using wordcount technique with Gensim library of Python.

V. Conclusion

This technique condenses the shorter version of e-news while retaining its overall meaning. The text Rank algorithm works in efficient way when used with Python Gensim library. It's used for Marathi language prominently. Summary consists of high ranked sentences which are combined together after ranking. The output summary will consist of the most representative sentences and will be returned as a string, divided by newlines. Wordcount parameter determines how many words will the output contain. The proposed method gives the output summary with specified number of words. This is a very precise method which is specially used for Marathi text, as the techniques and tools perform in a different way with different languages.

References

[1]. Apurva D. Dhawale, Sonali B. Kulkarni, and Vaishali Kumbhakarna, "Survey of Progressive Era of Text Summarization for Indian and Foreign Languages Using Natural Language Processing", ICIDCA 2019, LNDECT 46, pp. 654–662, Springer Nature Switzerland AG, 2020.

[2]. Apurva D. Dhawale, Sonali B. Kulkarni, and Vaishali Kumbhakarna, "A Survey of Distinctive prominence of Automatic Text Summarization Techniques Using Natural Language Processing", International conference on mobile computing and sustainable informatics, (ICMCSI), Springer conference, Tribhuvan university, Nepal, 2020.

[3]. Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna, "Automatic Pre-Processing of Marathi Text for Summarization", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Volume-10 Issue-1, October 2020.

[4]. Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna, "A Machine Learning Approach for Automatic Unsupervised Extractive Summarization of Marathi Text", International Journal of Creative Research Thoughts (IJCRT), ISSN: 2320-2882, Volume 8, Issue 11, November 2020.

[5]. Pooja Bolaj, SharvariGovilkar, "Text Classification for Marathi Documents using Supervised Learning Methods", International Journal of Computer Applications (0975 – 8887), Volume 155 – No 8, December 2016.

[6]. Deepali K. Gaikwad, Deepali Sawane and C. Namrata Mahender, "Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 51-54, 2018.

[7]. Yogeshwari V. Rathod,"Extractive Text Summarization of Marathi News Articles", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 07,July 2018.

[8]. Shraddha A. Narhari, RajashreeShedge, "Text Categorization of Marathi Documents using Modified LINGO", IEEE, 2017

[9]. Jaydeep Jalindar Patil, Prof. NagarajuBogiri, "Automatic Text Categorization-Marathi documents", International Conference on Energy Systems and Applications (ICESA 2015), IEEE, 2015.

[10]. Prakhar Sethi, Sameer Sonawane, SaumitraKhanwalker, R. B. Keskar, "Automatic Text Summarization of News Articles", International Conference on Big Data, IoT and Data Science (BID) Vishwakarma Institute of Technology, Pune, Dec 20-22, IEEE, 2017

[11]. N. Dangre, A. Bodke, A. Date, S. Rungta, S.S. Pathak, "System for Marathi news clustering", 2nd International conference on Intelligent computing,communication & convergence, bhubaneshwar, ELSEVIER, 2016.

[12]. Mr. Shubham Bhosale, Ms. Diksha Joshi, Ms. VrushaliBhise, Prof.Rushali A. Deshmukh, "Marathi e-Newspaper Text Summarization Using Automatic Keyword Extraction Technique", International Journal of Advance Engineering and Research Development Volume 5, Issue 03, March -2018.

[13]. Anishka Chaudhari1, Akash Dole2, Deepali Kadam3, "Marathi text summarization using neural networks", International Journal of Advance Research and Development, Volume 4, Issue 11, 2019.

[14]. Wan, X., Luo, F., Sun, X., Huang, S., Yao, J.-G.: Cross-Language Document Summarization via Extraction and Ranking of Multiple Summaries. Springer, London (2018). Part of SPRINGER Nature.

[15]. Liu, X., Webster, J.J., Kit, C.: An extractive text summarizer based on significant words. In: ICCPOL 2009 LNAI, vol. 5459, pp. 168–178. Springer, Berlin/Heidelberg (2009)

[16]. http://www.cdacnoida.in/snlp/digital_library/text_summ.asp

[17]. Gupta, V., Lehal, G.S.: Automatic Punjabi text extractive summarization system. In: International Conference on Computational Linguistics, COLING 2012, pp. 191–198. IIT, Bombay (2012)

[18]. Babar, S.A., Patil, P.D.: Improving performance of text summarization. In: International

[19]. Conference on Information and Communication Technologies, Procedia Computer Science, vol. 46, pp. 354–363. Elsevier (2014)

[20]. Vaishali V. Sarwadnya, Sheetal S. Sonawane, "Marathi Extractive Text Summarizer using Graph Based Model", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE, 2018.

[21]. Dr. Rajesh. S. Prasad, Uplavikar Nitish M, Wakhare Sanket S, "Feature Based Text Summarization", International journal of advances in computing and information researches, pune, January 2012.

[22]. Nilesh R. Patil,Girish Kumar Patnaik, "Automatic Text Summarization with Cohesion

[23]. Features", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 8 (2), 194-198, ISSN: 0975-9646, 2017