Comparing the Analytical Algorithms for Unsupervised e-News Summarization Using Machine Learning Tactics

Apurva D. Dhawale(✉), Sonali B. Kulkarni, and Vaishali M. Kumbhakarna

Department of CS & IT, Dr. B. A. M. University, Aurangabad, Maharashtra, India
addhawale@gmail.com

Abstract. The Text mining domain is acquiring importance and is used everywhere as the electronic documents are flooding the internet now a days. One of the significant e-content used by the students giving competitive exams is Marathi e-news articles which can be summarized in an extractive way without changing its meaning. We are trying to develop a system which will help students to extract current information from e-news in a reduced amount of amount of time. The comparison of the two recent and advanced methods is achieved for accentuating on the better one. The extractive summarization is performed using 3 different methods of TextRank algorithm i.e., by using summarize function, by using ratio and by using wordcount. This paper emphases on the analysis of results we got from TextRank algorithm using Gensim Framework with ROUGE method.

Keywords: ROUGE · TextRank algorithm · Gensim · indic · iNLTK · Marathi

1 Introduction

The breaking study area now is that Natural Language Processing (NLP) techniques which has not got the leading attention as of other domains. The major research is found in the English language. So, it's a challenge to use and build NLP applications in vernacular languages [5].

Many practical approaches have been covered by NLP which is shown in Fig. 1 [2]:

Finding the patterns in text documents is a preliminary task of Natural Language Processing and Machine Learning domain. So, it needs to have a machine and human communication which will inspire the users for initiatives and for better text mining or summarization practices. Figure 2 represents the types of summarizations one can choose for the respective job [7]:

The importance of saving time by summarizing text is increasing rapidly and the amount of e data is also growing in the high proportion. People tend to have exact meaningful data with minimum length of text document and hence, ATS (Automatic Text Summarization) is playing an significant role in the language barriers as they prefer to read text in their own regional or spoken language.

In Maharashtra, Marathi language is much spoken and treated as its regional language. In the literature study it has been found that, Text summarization of Marathi

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 141–148, 2023.
https://doi.org/10.2991/978-94-6463-136-4_15

Fig. 1. A Real time approached of NLP to solve problems

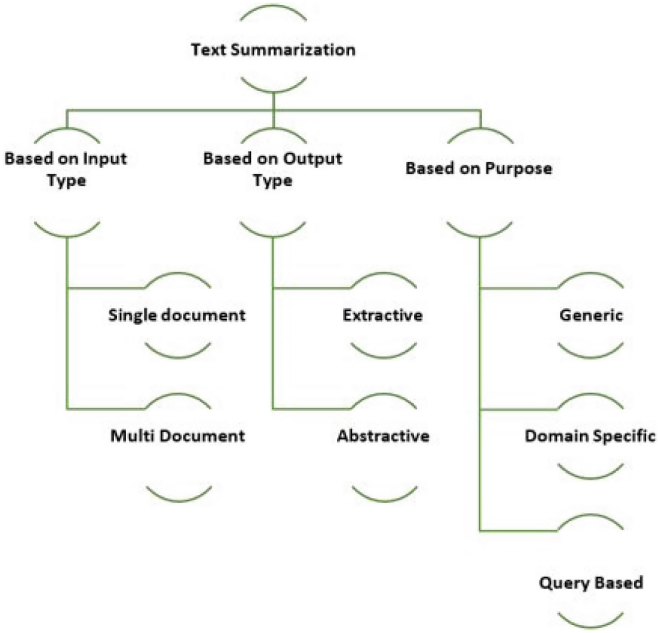

Fig. 2. The Categories of Text summarization

language has no pragmatic influential tool, or system that gives high efficacy in summarizing Marathi text. So, this research prominently focuses on an extractive based approach using Text ranking algorithm. Here, the first step is to read the input text document, find its length, and apply some basic pre-processing steps to finally derive important sentences and prepare a meaningful summary out of it.

There are 2 types of learning: 1) Supervised and 2) Unsupervised (Fig. 3).

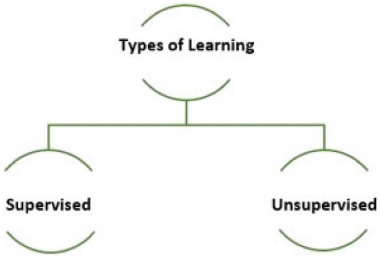

Fig. 3. Types of learning

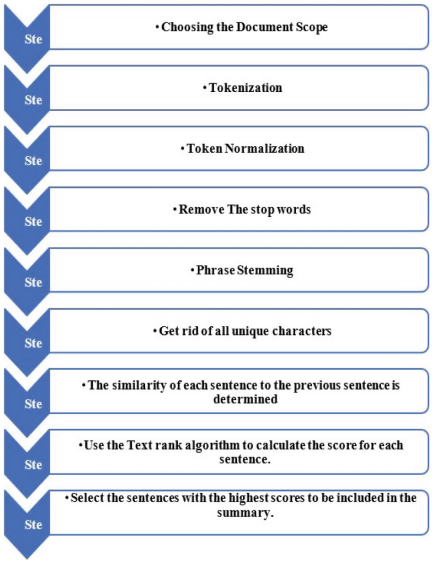

Fig. 4. Steps for generating summary using TextRank algorithm

1.1 TextRank Algorithm

The TextRank is a machine learning algorithm which has given an exponential growth in the results of Marathi Text summarization. For pre-processing of input text TextRank algorithm is used which determines the correlated sentences and keywords in a specified text. Those sentences are combined together to generate the summary of input text. As it is a graph-based method, the significance of a vertex is obtained based on the whole information provided by the graph [2].

It works with unsupervised graph-based ranking model for summary generation and follows following steps for generating summary from input text (Fig. 4).

1.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Algorithm

To aid the optimization of nondifferentiable functions such as ROUGE, there are various Reinforcement Learning (RL) methods [1] and the ROUGE signifies data as an information table where data has determinate attributes.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) support in logically deciding the nature of a summary by comparation with human (reference) synopses considered typically as the ground truth. Various sentences are compared by using the different categories of ROUGEs. The ROUGE-L, ROUGE-N, and ROUGE-S decides the granularity of texts which is the comparison between the reference summaries and system summaries [2].

The ROUGE was developed as an assessment measure which resolute how well an automatic summary covered the matter, present in an original text [8].

First step is to find out the predicted value of each sentence similarity score (y), then the prioritization based on y, of n top ranking sentences are considered for the peer summary with respect to limits given for summary length, & at last ROUGE is used to find the relevant summaries. Here we evaluate the performance of this model [9].

The importance of the procedures can be proven by comparing the human and automated summaries [10]. ROUGE-N count the number of overlapping units of n-gram, ROUGE-L count the word sequences and ROUGE-S count the word pairs between the candidate summary and the reference summaries.

When using ROUGE, the following considerations are important:

- Multiple reference summaries
- Pre-processing configurations
- Other configurations
- Comparing different systems [11].

A set of distinct ways to quantify the quality of a system summary is called as ROUGE metric [12].

2 Proposed Methodology

This study indicates the growing importance of e-text and its summary for a worthy reduction of human efforts with time. Eventually, researchers are summarizing the text in their own regional language for ease of use. There are various domains where text summarization will play a vital role like academics, banking, shares, research, news, and many more. Here we have focused on e news summarization of Marathi text which will be helpful for the students who are preparing for MPSC, UPSC examination. They do need daily updates and clusters of e-news.

The text document need to be classified and features are then put into combinations so that it can be summarized. We have used two datasets for our study: 1. TextSummarizer Marathi Dataset 2. Marathi news documents dataset master.

The first dataset contains 1135 Marathi e-news articles from 5 diverse domains like film industry, banking, polity, general knowledge and sports; the second one contains 100 mixed news articles of all domains.

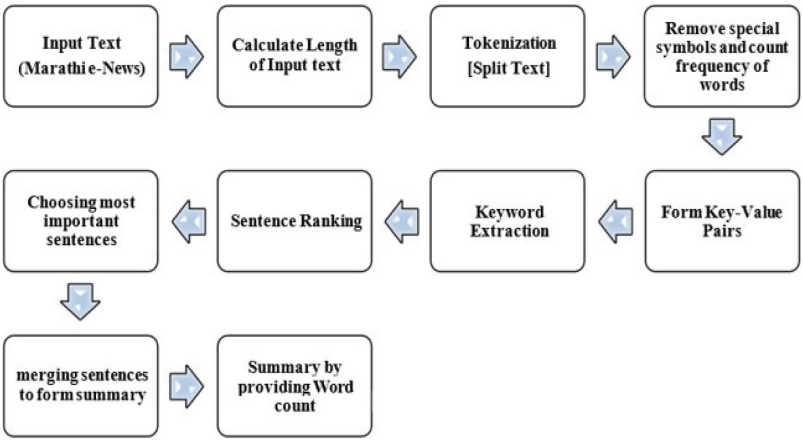

Fig. 5. Proposed Methodology

There are various python NLP libraries one should know while dealing with indic text. They are: Natural Language Toolkit (NLTK), Gensim, Polyglot, TextBlob, CoreNLP, Spacy, Pattern, Vocabulary, Quepy, iNLTK (Natural Language Toolkit for Indic Languages), Indic NLP Library [6].

There are various packages in the Python suitable for this sort of task. This research uses some of them. Primarily used packages are Gensim, PyPI, Scipy, Conda, Numpy, & Corpora. The NumPy and SciPy pack-ages are installed first and then we used Gensim for scientific computing. The Gensim is developed by GitHub and is commercially supported package. The Corpora is also substantial package for summarizing text.

In this research we have used Gensim framework on the test samples from first dataset where the average result of accurate summary generation is 96.66%, and the same framework is applied on the second dataset to have comparative analysis between two datasets (Fig. 5).

The memory independent algorithms are used by Gensim, which makes it tremendously effective in terms of memory usage and processing speed. Most glaringly, Gensim is robust and scalable as well [6]. It can handle large data or text by using data streaming and incremental algorithms available online.

Firstly, Marathi text is given as input to the system and its total length is determined. This length is compared with the text which is summarized at the end, to get the ratio to be tested.

The word split step is accomplished by using mytext.split() function & it is saved in a variable. Sentences are split into words, these words are then forwarded to count the frequency, it is then stored in an empty array. To find the frequency count, the get () function is used and this counter helps to get exact count of individual word identified from sentences.

The efficacy of an algorithm is dependent on the language used for deriving summary. In the subsequent step Key Value pairs are created. This system supports to improve the results by pre-processing data, and providing improved usefulness of Marathi text.

Gensim library of Python includes 3 methods for Marathi ATS, which helps user to produce meaningful summary of input Marathi e-news article. This algorithm provisions unsupervised extractive Marathi text summarization which generates summary by combining top rated sentences from the input data file.

The first method in Gensim TextRank algorithm is to use the Summarize() function from, the summary of input text by using 0.2% ratio. i.e. 20% is delivered by using summary. This needs at least 200 characters as an input text.

The second method is wordcount, which has to be given as an input from user. We have considered 150 words with which user can have meaningful summary. Therefore, the summary is generated with 150 words. This value of count may be increased or decreased as per user necessities.

The Third Method takes ratio value which ranges from 0 to1. User needs to give value in between 0 and 1. Here, 0.3 refers to 30% of summary, similarly 0.5 refers to 50% summary which we have considered for our testing news articles. The default value is 0.2 [4].

3 Comparative Evaluation of Algorithms

The Marathi news document dataset master has the summaries generated using ROUGE method. The result of ROUGE method is kept as a standard for comparative analysis between the two methods. The resultant summary generation is represented in Fig. 6. This blue bar represents the length of original text in the e-news article, orange bar represents the summary generated by using TextRank algorithm and the Gray bar represents the summaries generated by ROUGE method. Figure 7 shows the analysis of two methods on tested sample e-news articles.

Here, the Blue bar represents the summary generated by ROUGE method and the orange line graph represents the summary generated by Summarize function in TextRank algorithm.

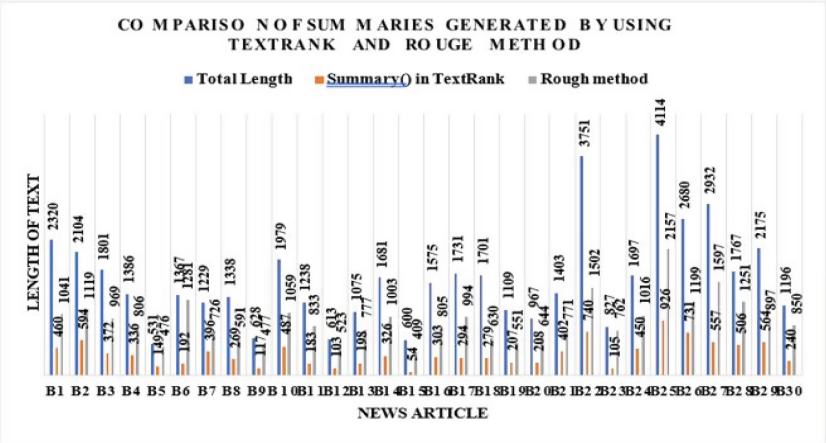

Fig. 6. Comparison of summaries generated by using TextRank and ROUGE method (with original text length)

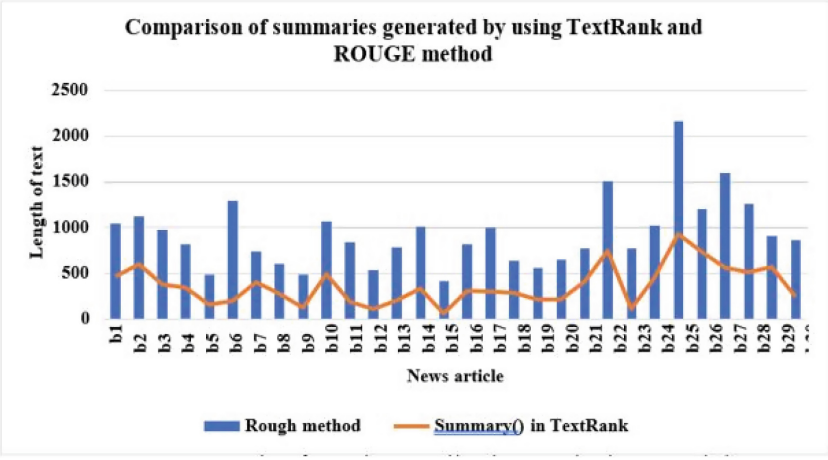

Fig. 7. Comparison of summaries generated by using TextRank and ROUGE method

4 Conclusion

The study of e news summarization using TextRank Unsupervised Machine Learning Algorithm gives a worthy result of summary compared with the ROUGE method. We used 3 methods for summarizing text using Gensim library of Python, which are summary (), summary by using ratio, & summary by using wordcount. The average result by combining all the 3 methods is 96.66% which is more than ROUGE method. So, this technique will prove to be precise and beneficial in terms of summarizing text to make effortless data abstraction for time saving.

References

1. Pradeepika Verma, Anshul Verma, "A Review on Text Summarization Techniques", Journal of Scientific Research Institute of Science, Banaras Hindu Univer-sity, Varanasi, Volume 64, Issue 1, India, 2020.
2. https://www.kdnuggets.com/2018/10/main-approaches-natural-language-processing-tasks.html, 2018
3. Divakar Yadav, Naman Lalit, Riya Kaushik, Yogendra Singh, Mohit, Dinesh, Arun Kr. Yadav, Kishor V. Bhadane,2022
4. Adarsh Kumar and Baseem Khan, "Qualitative Analysis of Text Summarization Techniques and Its Applications in Health Domain", Hindawi Computational Intelligence and Neuroscience Volume, Article ID 3411881, 14 pages. https://doi.org/10.1155/2022/3411881, 2022
5. 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), Volume 8, Issue 11 | ISSN: 2320-2882, November 2020.
6. https://www.analyticsvidhya.com/blog/2020/01/3-important-nlp-libraries-indian-languages-python/

7. https://machinelearningknowledge.ai/amazing-python-nlp-libraries-you-should-know/#Gensim
8. https://devopedia.org/text-summarization
9. Meetkumar Patel, Adwaita Chokshi, Satyadev Vyas, Khushbu Maurya, "Machine Learning Approach for Automatic Text Summarization Using Neural Net-works", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 7, Issue 1, January 2018
10. Alexander Dlikman and Mark Last, "Using Machine Learning Methods and Linguistic Features in Single-Document Extractive Summarization", Proceedings of DMNLP, Workshop at ECML/PKDD, Riva del Garda, Italy, 2016.
11. Farshad Kiyoumarsi, "Evaluation of Automatic Text Summarizations Based On Human Summaries", 2nd Global conference on linguistics and foreign language teaching, LINELT-2014, ELSEVIER, ScienceDirect, Dubai – United Arab Emirates, December 11–13, 2014.
12. Wenyi Tay, Aditya Joshi, Xiuzhen Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-faced ROUGE: Examining the Suitability of ROUGE for Opinion Summary Evaluation. In Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, pages 52–60, Sydney, Australia. Australasian Language Technology Association. 2019.
13. Marcello Barbella, Genoveffa Tortora, "ROUGE metric evaluation for Text Summarization techniques", Expert Systems with Applications May 16, 2022