International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

A Survey of Extractive Text Summarization for Regional Language Marathi

Deepali Kadam

Asst. Prof, DMCE, Maharashtra, India

Abstract - *An Automatic text summarization is a data reduction process to exclude unnecessary details and present important information in a shorter version. Technology of automatic text summarization plays an important role in information retrieval and text classification, and may provide a solution to the information overload problem we are facing due to this ocean of data. Though there are major basic two ways of summarizing data: abstractive and extractive, we are going to focus on later one. One way to summarize document is by extracting important sentences in the document and that is what we call as extractive summarization. Though the technique has been getting used since almost seven decades now, quite a less research work has been done in Marathi language considering summarization. As per the latest survey (2020 & 2021), Marathi holds 15th rank globally in terms of having most native speakers. And thus, in this paper, the survey for the work done in extractive summarization of Marathi language, though it is limited one, has been presented and on the basis of that the gaps where the improvement can be done for the better results have been listed out.*

Key Words: **Marathi Text summarization, Indian language, Extractive summarization, Regional language, Automatic summary, feature extraction.**

1. INTRODUCTION

Knowledge is wealth and so is time. There is a huge amount of data available in this massive empire ruled by the internet today. People are too busy to read the wordy documents, but they prefer to read concise ones. And thus, to serve the society with automatic ready gist of what they aspire to read without wasting their valuable time, text Summarization comes to the rescue.

Text summarization provides the user with condensed description of documents and a non redundant presentation of facts found in the document. Automatic text summarization has been in existence since last few decades. While there are a number of problems remaining to be solved, the field has seen quite a lot of progress, especially in the last two decades, on extraction-based methods. The rapidly growing popularity of Internet has become an important symbol of the information age. In the phase of the flood tide of electronic literature, to search for a way to read the necessary and compact information, is clearly inefficient and infeasible. Therefore, by reading the summary to obtain useful information is the best way to save our time.

Automatic Text Summarization plays an inevitable role in everyday life. For example, headlines of news, summary of technical paper, review of book or preview of a movie. There are two major basic types: Abstractive and Extractive. Extractive process is significantly different from human based text summarization i.e. abstractive one. Since human can capture and relate deep meanings and themes of text documents while automation of such a skill is very difficult to implement.

One of the obvious questions to ask in doing summarization is "what are the properties of text that should be represented or kept in a summary?" Summarization generally happens in two phases: Pre-processing and Processing. Pre-processing further subdivides into sentence segmentation, tokenization, stop word removal and stemming. The most common features for pre-processing, researchers have considered till now are average tf-isf, sentence length, sentence position, numerical data, sentence to sentence similarity, title word feature, thematic word feature, proper noun feature etc.

Language is surely a tool we use to communicate and express ourselves, a means towards an end, if you will. However, one's mother tongue is way more than being merely a tool. Marathi language is ranked 15th in terms of having 83 million native speakers in India alone, not to speak of the overall 95 million speakers who speak Marathi as their mother tongue. It is a regional and official language of Maharashtra. Even then nominal work has been done in Marathi Text Summarization. So, Marathi language has been chosen as language of study. Marathi is written in the Devanagari script which has one of the largest alphabet set.

2. LITERATURE SURVEY

2.1. A Survey of Automatic Text Summarization System for Different Regional Language in India

Virat Giri and Dr. M. M. Math[9] contemplated automatic text summarization system for different regional languages in India. They have done most of their work from the ground up. This is because, a rare work had been done before 2016 in Marathi language processing. They studied the techniques in various languages and tried to apply it for Maharashtra's official language. They developed Marathi stemmer, Marathi proper name list, English-Marathi noun list, Marathi keywords extraction; Marathi rule based named entity recognition etc. These lexical resources are used in pre-processing and processing steps. The three major steps

were Pre processing, stemming, sentence ranking and summary generation. For summary evaluation, Intrinsic and Extrinsic measures have been suggested.

2.2. Rule based Stemmer using Marathi WordNet for Marathi Language

Pooja Pandey and Dhiraj Amin[8] developed stemmer defining their own rules for stemming for Marathi language. They used Marathi wordnet for the same. Two modules were presented. First is preprocessing module and second is stemming module. The stemming module consists of root verification, suffix removal and inflection removal. They extended the rule based approach by involving Name entity. They created their own stem exception dataset. They tried to handle over-stemming and under-stemming issue in their work and succeeded reducing the error proportionally. There is still a lot more scope for improving the stemming result as hardly any work has been done in Marathi language considering the complexity in its word formation. So we can define more rules in future having more detailed analysis of morphological study of Marathi word formation.

2.3. MarS: A Rule-based Stemmer for Morphologically Rich Language Marathi

Harshali Patil and Ajay Patil[7] described the stemmer for Marathi language. They defined rules based on longest common match principle. Stemming is an important step while preprocessing any text for summarization or information retrieval. Marathi language is highly inflected language. Dealing with morphological variations in Marathi language is a tough job. Yet approximately 79.97% accuracy is achieved when tested on a collection of 4500 unique words from the news corpus among nine runs.

In Marathi language suffixes(प्रत्यय) are mostly applied on inflected form (samanya roop : सामान्यरूप) and not on original form(मूळ रूप). Thus stemming becomes more difficult for Marathi language. They studied the orthographic structure (शुद्धलेखन विषयक रचना) for Marathi language and defined the generic rule set for stemming taking into consideration the important points in formation of words like suffixes, inflections and compilation of suffix list. Though the accuracy was quite satisfactory, they fell short to handle over stemming and under stemming error issues in stemming.

2.4. Marathi e newspaper text summarization using automatic keyword extraction technique

Mr. Shubham Bhosale along with Diksha Joshi[6] used keyword extraction technique for Marathi e newspapers. They preferred ranking keywords rather than ranking sentences in the original text. The size of the summary is predefined and fixed i.e. thirty to forty percent of the news article, user is desired to read. Less complex statistical approach has been chosen for selecting keywords. Though it is less complex, comparatively it gives better performance. By using these highest ranked words, sentences consisting of these words are picked to form a summary.

2.5. Extractive Text Summarization of Marathi News Articles

Around three years ago, Yogeshwari Rathod[5], has auspiciously protracted the traditional Text-rank algorithm for graph based ranking model built for news articles in Marathi language. They have brought forward two inventive unsupervised methods for keyword and sentence extraction. The graph is designed analogous to text and graph vertices are treated as ranking units. They aimed to rank entire sentences. Thus, for each sentence in the text, there will be an equivalent vertex. A specific relation has been designated that determines an interdependence between two sentences if there is a similarity relation between them. Similarity is measured as a function of their content overlap. Also, for evaluation purpose, 2 summaries are created by human and compared them with the automatic generated summaries using ROUGE tool.

2.6. Marathi Extractive Text Summarizer Using Graph Based Model

Vaishali V. Sarwadnya and Sheetal Sonawane[4] developed graph based automatic text summarizer for Marathi language. For the evaluation purpose, ROUGE-N metric was used where N varies from 1-4. They listed out challenges we need to face while working with such a morphologically rich language. Less availability of Dataset is one of the major issue when it comes to Marathi Language. Deficiency of ready-to-use data for test and training, creating your own dataset itself becomes one tedious task. For their own work, they went for EMILLE (Enabling Minority Language Engineering) corpora. Also they threw light on ambiguity, dialects and variations in spelling or formation of words. As the model is graph based, they had different options like bushy path, aggregate similarity algorithm, Text Rank algorithm among which they chose the last one. They experimented with features like thematic similarity and positional distribution of sentence. Their results were better than the others and can be improved if we add more features like semantic ranking.

2.7. Multi-Document Text Summarization of Marathi Regional Language

Around two years ago, Sujata Lungare and Aman Jain[3] made use of web application for multiple articles in Marathi language. They preferred probabilistic model for summarizing different source documents into only lone summary document. They focused on evaluation of summary. For solving the evaluation purpose, they mainly worked on precision, recall, F-score compression ratio and retention ratio parameters. The limelight of Probabilistic algorithm is the probability of words used in merged document.

Compression ratio is not fixed but user friendly. A per user requirement, size of summary is customized. They used Markov Chain Model and Bag of Words parser. The output was quite relevant and close to desired one. Though it is a web service, we can enhance it on cloud or mobile platform considering broad spectrum.

2.8. Automatic Pre-Processing of Marathi Text for Summarization

Apurva D. Dhawale and Sonali B. Kulkarni[2] have explored the pre processing techniques for Marathi news articles in the last year.As per their observation, LINGO [Label Induction Grouping] algorithm can be used for improving results efficiently in Marathi document summarization. They noticed that the ways suitable for Indian Languages are Scoring of sentences, ROUGE evaluation, Language-Neutral Syntax (LNS), Support Vector Machine : SVM classifier, Hybrid Algorithm. Marathi being an Indian language, these methods applies to Marathi too. They have focused on Marathi text processing using Text raking, clustering, lexical chain, domain specific summarization algorithms. They have used python library to work with NLP and Information Retrieval. They have formed the Key Value pair giving a list of words appending with its frequency count. They tried to use this word based count for creating more capable summarization system.

2.9. Marathi Text Summarization for News Articles using Sequence To Sequence with Attention Mechanism

Kavya Nair, Rushali Deshmukh[1] used sequence to sequence model along with attention mechanism for text summarization in Marathi language last year only in December 2020[1]. They created their own dataset as collection of News Articles. They have used encoder-decoder LSTM to improve the quality of the generated summaries. They proposed the first model of extractive summarization using text rank system, but it failed due to grammatical inconsistency. Then they went for abstractive model using encoder-decoder. They realized that it only works for short sequences as encoder finds it difficult to memorize long sequences. To overcome this issue, they went for attention mechanism in their next model achieving an accuracy of 86-88%. While dealing with abstractive, they comprehended how difficult it is to decipher Marathi words for the summary. Going ahead they introduced LSTM Long Short Term Memory at both the side encoder and decoder to finally achieve accuracy of 76%. There is a scope of improvement when it comes to avoid repetitive words in summary and interactive GUI.

3. Proposed Model

Flow of our proposed system is as follows:

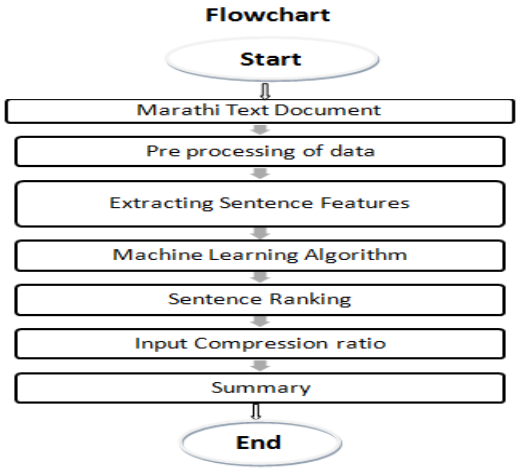

Fig -1: Flow of proposed system.

Extractive text summarization process can be divided into two steps: Pre Processing step and Processing step. The brief of sub steps are shown below.

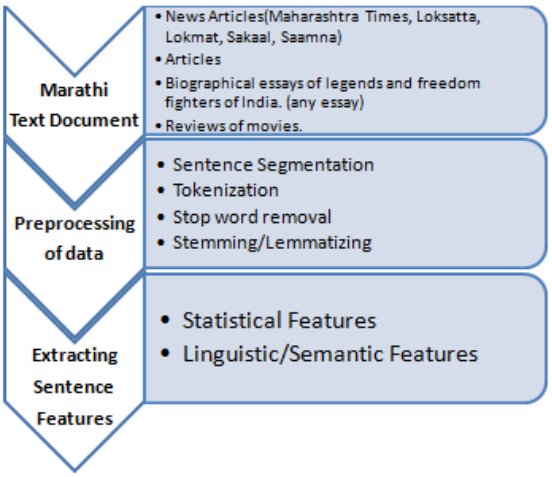

Fig -2: A brief of Pre-processing and Processing data

3.1. Preprocessing step:

Preprocessing step includes:
1) Sentence segmentation: Basic unit that possibly conveys independent meanings is detected in a form of individual sentence [3]. In Marathi language, sentence is segmented by identifying boundary of sentence that ends with purnaviram (.).

2) Tokenization : The sentences are further divided into discrete bits or tokens(words). It excludes special characters, such as punctuation, spaces and special symbols between words. Punctuations (विराम चिन्ह) in Marathi

language consists of पूर्ण विराम (.), उपविराम (:), अर्ध विराम (;), etc.

विरामचिन्हे व त्याचे प्रकार

प्रकार	चिन्ह
पूर्णविराम	(.)
स्वल्प विराम	(,)
अर्धविराम	(;)
अपूर्णविराम	(:)
प्रश्नचिन्ह	(?)
उद्‌गारवाचक चिन्ह	(!)
अवतरण चिन्ह	(“ ”) (‘ ’)
संयोगचिन्ह	(-)
अपसरण चिन्ह	(-)

Fig -3: Examples of Punctuations in Marathi Language

3) Stop Word Removal: Stop Words include function words, articles, prepositions, conjunctions, prefix, postfix, etc. i.e. common words that carry less important meaning than keywords. Hence should be eliminated.

1. क्रियाविशेषण अव्यय व त्याचे प्रकार	
1.1.कालवाचक क्रियाविशेषण अव्यय	
अ. कालदर्शक -	उदा. आधी, आता, सद्य, तूर्त, हल्ली, काल, उद्या, परवा, लगेच, केव्हा, जेव्हा, पूर्वी, मागे, राजी, दिवसा इत्यादी.
ब. सातत्यदर्शक -	उदा. नित्य, सदा, सर्वदा, नेहमी, दिवसभर, आजकाल, अद्याप इत्यादी.
क. आवृत्तीदर्शक -	उदा. फिरून, वारंवार दररोज, पुन्हा पुन्हा, सालोसाल, क्षणोक्षणी इ.
1.2. स्थलवाचक क्रियाविशेषण अव्यय :	
अ. स्थितीदर्शक -	उदा. येथे, तेथे, जेथे, वर, खाली, कोठे, मध्ये, अलीकडे, मागे, पुढे, जिकडे-तिकडे, सभोवताल इत्यादी
ब. गतिदर्शक -	उदा. इकडून, तिकडून, मागून, पुढून, वरून, खालुन, लांबून, दुरून
1.3. रितीवाचक क्रियाविशेषण अव्यय -	असे, तसे जसे, कसे, उगीच, व्यर्थ, फुकट, आपोआप, मुद्दाम, जेवी, तेवी, हळू, सावकाश, जलद इत्यादी.
2. उभयान्वयी अव्यय	उदा. व, अन्, आणि आणखी, न, शि, शिवाय, अथवा, वा, की, किंवा, अगर , म्हणून, याकरिता, सबब, यास्वत, तेव्हा, तस्मात् इत्यादी.
3. शब्दयोगी अव्यय	साठी, कारणे, करिता, अथा, प्रीत्यर्थ, निमित्त, स्तव, शिवाय, खेरीज, विना, वाघून, व्यक्तिरिक्त इत्यादी
4. केवलप्रयोगी अव्यय	अहाहा, वाहवा, वा, ओहो, अबब, छेछे, अहं इ.

Fig -4: Examples of Stop words in Marathi Language

4) Stemming : Reducing the inflection to obtain the root word is called as a Stemming. In Marathi language there are many suffixes (प्रत्यय) as per their cases (विभक्ती). It splits the word into the root word or approaching towards root word and syntactically similar words, such as plurals, verbal variations, etc. e.g. walk, walking and walked are counted as same and derived from a stem word walk.

3.2. Processing step:

In processing step, we evaluate the value of features of each sentence. Weights are assigned to each sentence. Higher ranked sentences are extracted for summary. Feature Extraction: True analysis of the article for summarization starts with this step. Feature term values range between 0 to 1. Six statistical and two linguistic features are used as follows:

Statistical Features: 1. Average TF-ISF (Term Frequency Inverse Sentence Frequency) 2. Sentence Length 3. Numerical Data 4. Sentence Position 5. Thematic Word Feature 6. Title Word Feature
Linguistic Feature: 1. Sentence to Sentence Similarity 2. Proper Noun Feature

Applying the formulae and logic to each sentence considering above features, we will get the top most sentences as per the highest rank holders. Customizing the result as per the user’s choice of compression ratio, we will be picking up the 1st n number of sentences from highest rankers, as the summary of the article.

4. CONCLUSIONS

We have presented a survey on the text summarization in Marathi language. Though a lot of work is going on in other languages, comparatively less work has been done in Marathi language. Also, being an inflection wise complex language, we need to focus mainly on its preprocessing phase of summarization. There is hardly any substantial work found for stemmer developed for Marathi language. Also in feature extraction, there is a scope of adding more features in future. The system proposed in this paper is under development.

REFERENCES

[1] Kavya Nair, Rushali Deshmukh, “Marathi Text Summarization for News Articles using Sequence To Sequence with Attention Mechanism,” December 2020.

[2] Apurva D. Dhawale, Sonali B. Kulkarni, Vaishali M. Kumbhakarna, "Automatic Pre-Processing of Marathi Text for Summarization," October 2020.

[3] Sujata Lungare, Aman Jain, "Multi-Document Text Summarization of Marathi Regional Language," May 2019.

[4] Vaishali V. Sarwadnya, Sheetal Sonawane, "Marathi Extractive Text Summarizer Using Graph Based Model," August 2018.

[5] Yogeshwari V. Rathod, "Extractive Text Summarization of Marathi News Articles," July 2018.

[6] Mr. Shubham Bhosale, Ms. Diksha Joshi, Rushali Deshmukh, "Marathi e newspaper text summarization using automatic keyword extraction technique," May 2018.

[7] Harshali B. Patil, Ajay s. Patil, "MarS: A Rule-based Stemmer for Morphologically Rich Language Marathi," July 2017.

[8] Pooja Pandey, Dhiraj Amin, Sharvari Govilkar, "Rule based Stemmer using Marathi WordNet for Marathi Language," October 2016

[9] Virat V. Giri, Dr.M.M. Math and Dr.U.P. Kulkarni, "9. A Survey of Automatic Text Summarization System for Different Regional Language in India," October 2016.

BIOGRAPHY

Deepali Kadam
(Asst. Professor)
(IT, DMCE, Maharashtra, India)