Marathi Text Summarizer Using Deep Learning Model

Shruti Bhoir[1], Tanvi Hule[2], Deepali Kadam[3]

[1]BE Student, Information Technology, Datta Meghe College of Engineering, Airoli, Maharashtra
[2] BE Student, Information Technology, Datta Meghe College of Engineering, Airoli, Maharashtra
[3] Asst. Professor, Information Technology, Datta Meghe College of Engineering, Airoli, Maharashtra

---***---

Abstract – *Text summarization is a technique which converts the original text into short text by selecting important sentences without changing its original meaning . It is difficult for human to convert text manually. We present technique for extractive summarization of articles for Marathi language, in which it will consists of feature extraction, selecting important sentences, paragraphs etc. from the original document and catenating them into shorter form using deep learning model. In this system, we develop system in two stages. First stage is Summarization of Domain Specific Marathi article. In Second stage we will extend our model for generic article will be tested on various Marathi inputs. Such a summarization technique is known for English articles, and doing it for Marathi news is the novel part of the work.*

***Key Words*: Marathi article, Summarization, Extractive summary, Feature extraction ,Deep learning.**

1.INTRODUCTION

Automatic Marathi text summarization is technique of shortening the original text into shorter form which will give exact meaning of the original text. Summarization can be classified into two groups: extractive and abstractive summarization. Maximum of summarization systems are for English and other languages. For Marathi language, automatic text summarization systems are less. There is very less work done on Marathi summarization systems.

1.1 Problem Definition

The overall scope of the project is to have a deeper knowledge of the techniques in Machine Learning, Deep Learning and Data Analysis in order to generate concise summaries of long texts, which lets a user see a summary. The scope also involves understanding of why Machine Learning are successful at phrasing sentences and how they treat some input words more important than the others by assigning the appropriate weights, and have a better overview internally of how Machine Learning. We chose Marathi over all other languages because there are no projects or summarizer created on Marathi language yet. We present technique for extractive summarization of articles for Marathi language, in which it will consists of selecting important sentences, paragraphs etc. from the original document and creating a summary out of it.

2. PROPOSED SYSTEM

The summary produced by summarization system allows the user to easily understand the content of original documents without having to read each the whole document.

2.1 Abstractive and Extractive:

A. Abstractive: Abstractive summarization consists of understanding the source text by using linguistic method to interpret the text and expressing it in own language.

B. Extractive: Extractive summaries involve extracting relevant sentences from the source text in proper order. The important sentences are selected by applying statistical and language features to the input text.

2.2 Modules

The proposed Marathi text summarization method is extraction based.

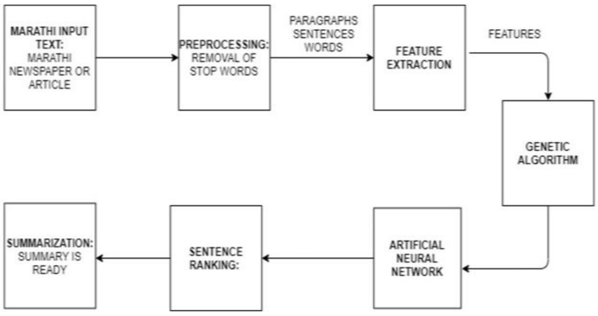
MARATHI INPUT TEXT: MARATHI NEWSPAPER OR ARTICLE
PREPROCESSING: REMOVAL OF STOP WORDS
PARAGRAPHS SENTENCES WORDS
FEATURE EXTRACTION
FEATURES
GENETIC ALGORITHM
ARTIFICIAL NEURAL NETWORK
SENTENCE RANKING:
SUMMARIZATION: SUMMARY IS READY

Fig -1: Design of Module

A. Pre-processing:

In pre-processing step the stop-word is removed, stemming and breaking the input document into a collection of sentences. The punctuation marks, characters like ; , ——: ()[]{} space character, tab space are removed. We will eliminate these words from text.

B. Stemming:

In stemming, a word is split into its stem and root word. A stemmer algorithm involves removing suffixes using a list of frequent suffixes. Pre-processing output is taken as input to stemming . It is further subdivided into two parts: Root verification, Suffix removal.

2.3 Feature Extraction

Actual analysis of the document to get summarized output start from here. Every sentence is represented by the feature terms vector and has a score based on the weight of feature terms.

a) Average TF-ISF(Term Frequency Inverse Sentence Frequency):It is used to evaluate how important a word is to a document in a collection.

$$TF=\left(\frac{word\,(term) occurence\ in\ sentence(Si)}{Total\ no.of\ words\ in\ sentence(si)}\right)$$

$$IF=\log\left(\frac{Total\ no.\ of\ sentences}{No.\ of\ sentences\ containg\ the\ term}\right)$$

$$\text{Avg TFISF(s)} = \sum TF * ISF$$

b) Sentence Length: It is used to eliminate the sentences which are too short or too long.

$$SL= \mathrm{Sin}((\ L - \ \mathrm{MinL}) * \left(\frac{\mathrm{Max}\theta - \mathrm{Min}\theta}{\mathrm{MaxL}-\mathrm{MinL}}\right)$$

c) Numerical Data: The numerical data is used to show the important mathematical or statistical analysis

$$ND=\left(\frac{No.\ of\ Numerical\ Data\ in\ a\ sentence}{Sentence\ Length}\right)$$

d) Sentence Position: The position of the sentence in the text, decides its importance.

$$SP= \mathrm{Cos}((CP- \mathrm{Min}\ V) *((\mathrm{Max}\theta - \mathrm{Min}\theta)/(\mathrm{MaxV}- \mathrm{MinV})))$$

$$New\_SP=[(1+Sp)/2]$$

e) Proper Noun Feature: It is used to select proper noun from the sentence as it has more importance.

$$PN=\frac{No.\ of\ Proper\ nouns\ in\ S}{\text{Sentence length of S}}$$

f) Thematic Word Feature: The most frequent content words are selected here.

$$TW=\frac{No.\ of\ Thematic\ nouns\ in\ S}{\text{Max(No. of Thematic Words)}}$$

g) Sentence to Sentence Similarity: It computes the similarity between each other sentence of the document, then add up those similarity values

$$ss=\sum_{j=1}^{N} \mathrm{Sim}(i,j)\ \ i \neq j\ \ and$$

$$Sim(i,j)=\frac{No.\ of\ words\ occured\ in\ Sentences(Sj)}{WT}$$

h) Title Word Feature: The words in the title carry higher weight and make the sentences containing them a possible candidate to be included in a summary.

$$TS=\frac{|(words\ in\ sentence)\cap(words\ in\ a\ title)|}{|\text{Total words in Title}|}$$

2.4 Feature Selection

For feature selection we are using Genetic Algorithm.
Genetic Algorithm:
Genetic Algorithm (GA) is a technique where search-based optimization is implemented based on Genetics and Natural Selection functionality. Usually it is used to find optimal or else near-optimal solutions for difficult problems which takes a longtime to solve. In short, it is a stochastic search algorithm which acts on population of possible solutions. Five phases are considered in a genetic algorithm. Initial population, Fitness function, Selection, Crossover, Mutation.

3. WORKFLOW

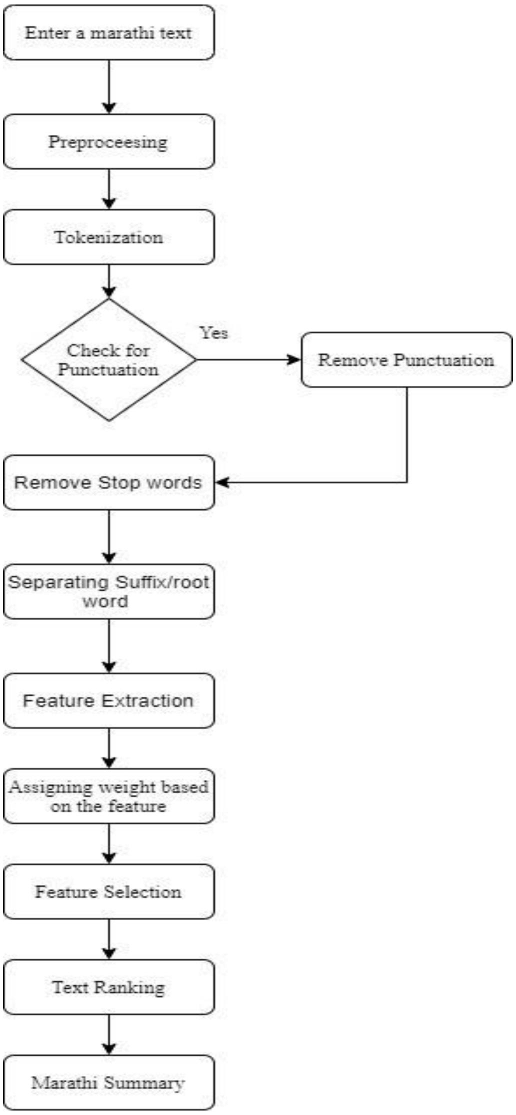

Fig -2: Workflow of Module

4. CONCLUSION

Marathi text summarizer is able to create an extractive summary using deep learning model. We have built the summarizer using python ,Natural Language Toolkit, Indic Natural Language Toolkit, Spacy .This system still has less accuracy which can be improved by having more effective feature selection algorithm. As there is very less work done on Marathi language our project will help many Marathi users to go through articles andsave time by just reading the summary. In future multiple languages can be proposed in a single system. The finalized output can be converted into audio format.

ACKNOWLEDGEMENT

Working on the project on “Marathi Text Summarizer” was a source of immense knowledge to us. However, this would not have been possible without the equal support of all members. I sincerely express my deep sense of gratitude to our Asst. Prof. Deepali Kadam, for her valuable guidance, continuous encouragement and support whenever required and timely help given to us throughout the course of this work

REFERENCES

[1] Anishka Chaudhari; Akash Dole; Deepali Kadam, “Marathi text summarization using neural networks”, Datta Meghe College of Engineering, Navi Mumbai, Maharashtra,2019.

[2] Mr. Shubham Bhosale; Ms. Diksha Joshi; Ms. Vrushali Bhise; Prof.Rushali A. Deshmukha, “Marathi e-Newspaper Text Summarization Using Automatic Keyword Extraction Technique”, RajarshiShahu College Of Engineering, Pune,2018.

[3] Yogeshwari V. Rathod, “Extractive Text Summarization of Marathi News Articles”, Department of Computer Science & Engineering, Vishwakarma Institute of Technology, Pune,2018.

[4] Arpita Sahoo; Dr.Ajit Kumar Nayak, “Review Paper on Extractive Text Summarization”, Department of Computer Science and Information Technology Institute of Technical Education and Research, S’O’A University, Bhubaneswar India,2018.

[5] Prof. Satish Kamble; Shivlila Mandage; Shubhangi Topale; Dipali Vagare; Prerana Babbar1, “Survey on Summarization Techniques and Existing Work” , PVG’s College of Engineering and Technology, Pune, Maharashtra, India,2017.

[6] Vipul Dalal, Latesh Malik, “Data Clustering Approach for Automatic Text Summarization of Hindi Documents using Particle Swarm Optimization and Semantic Graph”, Department of Computer Science & Engineering, Nagpur, Maharashtra, India,2017.

[7] Kanitha.D.K;D. Muhammad Noorul Mubarak, “An overview of extractive based automatic text summarization systems” Department of Computer Science, University of Kerala, Kariavattom, India,2016.

[8] Deepali P kadam; Mrs. Nita Patil; Mrs.Archana Gulathi,“A Comparative Study Of Hindi Text Summarization Techniques: Genetic Algorithm and Neural Network”,Department of CSE, Datta Meghe College of Engineering, Airoli, Navi Mumbai, India,2015.

BIOGRAPHIES

Shruti Bhoir
is student, Information Technology, Datta Meghe College of Engineering, Airoli, India.

Tanvi Hule
is student, Information Technology, Datta Meghe College of Engineering, Airoli, India.

Asst. Prof. Deepali Kadam
is Asst. Professor, Information Technology, Datta Meghe College of Engineering, Airoli, India.