The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

A TEXT SUMMARIZATION SYSTEM FOR MARATHI LANGUAGE

VAISHALI P. KADAM[1] SAMAH ALI ALAZANI [2] and C. NAMRATA MAHENDER[3]

[1, 2, 3]Department of Computer Science & I.T. Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra.
Email: [1]vaishu7817kadam@yahoo.in, [2]alazani183@gmail.com, [3]cnamrata.csit@bamu.ac.in

Abstract

Text summarization is the most popular application and a challenging task in the natural language processing. It is important for searching the specific information within the short time span from the input document. It is presently in demand to have quick information access as a summary to make a conclusion about the document text. This summary always presented with limited word and specific information contents for the search item. Summarizer systems are capable of generating a short version of the overall text after the analysis of the text it always retain its original meaning and the actual theme in the summary text. There are many automated summarizer systems developed for various Indian languages but still these systems are not achieved the matured stage. This paper proposed a methodology for development of the automated text summarization technique for Marathi language. We have got 44.48% compression accuracy for the summary by our system.

Keywords: Abstractive summarizer, Extractive summarizer, Linguistic analysis, Machine learning, Natural language generation, Natural language processing, Text summarization.

1. INTRODUCTION

In the present era there are unlimited sources available for getting information on a specific topic or a single search item. There are many language tools developed by researchers for various international and national languages. These information tools or language resources are natural language processing applications specifically designed in a particular language or for some limited languages. Languages are used by the people for communication and through which they perform exchange of knowledge and communicate in the society. Human languages are known as natural languages. Natural language processing or NLP is the sub-field of computer science and artificial iintelligence. There are many applications designed using natural language processing like information retrieval system, text summarization, automatic question answering system, language translators, sentiment analysis etc. Designing such system goes through five different phases of NLP such as lexical or structure analysis, parsing or syntax analysis, semantic analysis, discourse integration, pragmatic analysis through which specific language related task can be achieved. In technology enhancement language interpretation, understanding is generally used for finding the exact meaning of the language and also caused to provide assistance to the information retrieval services and their exchanges. It is most important for some tasks to provide the quick information. It is difficult to read the whole text and then search the specific information. An automated summarizer provides the solution to search this information in the form of summary. There are two main methods for getting or generating summary from the source document text. One the extractive and another is the abstractive text summarization method. Both the methods used to generate the summary but the main difference is that extraction methods only extracts or selects important sentences or text from the original source document. This type of summary

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

consist of same sentences present in the source document and also follows the same sequence of their appearance exactly in the same way they appeared in the source text. This is the main drawback of the extractive method that it does not follow own sense to generate the summary. But in case of abstractive summary machine uses an expert techniques to read the whole text from the input document and understand its proper meaning used by the speaker with its exact intention. [1] For this task input will be accepted in a particular language and summary generated in that same language. In this paper we proposed a summarizer which is used as a language tool developed for Marathi language.

1.1 Natural Language Processing

It is seen that there are many languages available on earth for different geographical areas. But a particular human language is used by or spoken by the people of some specific category or for a specific geographical area. Different types of natural languages are the main source for convergence and development of the society. Therefore new innovations and whole societal development is mostly based on these languages. In general, natural language or once mother tongue is the only source for changing the whole atmosphere and real development of the society. In this era of information technology a special branch of computer science and artificial intelligence performs study of such human languages known as the computational psycho-linguistics. It studies the human languages for machines to develop their abilities through the real time perception and exact make the machines trained to achieve equivalence in their language understanding and decision making. It causes machines to properly response to the people when these machines used for providing information services. Machines generate their responses in human languages based on their experience and training given to them. So a good language tool is the requirement in the society to serve them using the technology. This advancement of language cause to achieve the progress and changing the look of the whole economy, the strength and power of the nation. This technical ability is developed using some special machine learning schema developed for the language.

1.2 Text Summarization

Text summarization is a challenging research task in NLP which provides meaningful summary of any given input document. Actually, it is the process of conversion of a lengthy text document to its shorter version without changing its meaning with overall original sense. This short version of meaningful text is called summary. This summary text is always less in size than half of the main text. The main task is the extraction or selection of important contents from the source text and then representation of text to generate a summary. Google summarizer is a good example of a text summarizer. This system helps users to find the important contents that is of most interested to them. The summarization system automatically accepts inputs either in the form of a single document or multiple documents or a query as an input and produces an abstract or an extract summary using various tools and techniques. This summary may also be indicative or informative as per the user requirements. [2]

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

1.2.1 Extractive Summarization

In this type of the text summarization, techniques are followed to select the most important sentences and paragraphs from the body of the text. This method follows mainly statistical analysis to rank the sentences for finding their relevance and importance in the document. After finding highly ranked sentences or their regions from entire document, this extracted sentences or extracts taken from the document can be re-ranked by combing and arranged them based on the similarity measures into a summary. [3]

1.2.2 Abstractive Summarization

In this abstractive summarization, generation of summary for the text is based on the understanding and regenerating the skill of the system to its short form. It is basically of two main types one is the structured based and another the semantic based approach. In this process each sentence is interpreted for prediction of its meaning based on overall language analysis. For this different methods like ontology, rule-based, hybrid, tree based, template based and lead and body phrase method etc. for generating the new summary. Abstraction involves paraphrasing the corpus using new sentences. Abstractive summary generation rewrites the entire document by building internal semantic representation, and then a summary is created using natural language processing. The abstractive summarization research works in Indian languages are in premature state when compared to other languages like English. French, Arabic, Spanish, Chinese, German etc. This is mainly due to the diversity in Indian languages and the lack of resources such as raw data, various NLP tools etc. [4].

1.3 Importance of text Summarizer in Marathi

Languages in India can be divided into Indo-Aryan languages and Dravidian languages. Indo-Aryan languages include Hindi-Urdu, Assamese, Bengali, Gujarati, Marathi, Punjabi, Rajasthani, Sindhi, and Oriya etc. Dravidian languages include languages like Malayalam, Tamil, Telugu, and Kannada etc. Though Malayalam and Telugu are Dravidian in origin, over eighty percent of their lexicon is borrowed from Sanskrit [5]. Human languages are spoken by the native people of some specific category or for a specific geographical area such natural languages are the main source for convergence and development of the society. Therefore new innovations and whole societal development is mostly based on these languages. In general, natural language or someone's mother tongue is the only source for changing the whole atmosphere and real development of the society. So a good language tool is the requirement in the society to serve them using the technology. This advancement of language cause to achieve the progress and changing the look of the whole economy, the strength and power of the nation. This technical ability is developed using some special machine learning schema developed for the language. An automatic summarizer is the need of technologies that do all the sorting and quickly identify the relevant information by its own and can generate the summary.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

2. LITERATURE SURVEY

Past literature study always guides the new researchers and innovative task to proceed in the right direction. This study helps to improve the current results and the overall task. In early researches on summarization most of the researchers concentrated mainly on extraction of important sentences instead of generation of own text as summary. There are many method for summarization that are based on statistical features of the sentences but they caused to produce extractive summaries than the abstractive summary. Below table 2.1 shows the previous task that has contributed in the development of summarization systems.

Table 1: Literature survey of existing summarization systems in Indian languages

Author	Language	Method or Approach	Achievement or System Response	Lacuna
Nikita Munot, et al. , 2014 [6]	Marathi	statistical and linguistic	Different methods for text summarization discussed with advantages and disadvantages	Need of abstractive summarizer to produce semantically related summary which is difficult to produce.
Sunitha et al. 2016 [7]	Marathi	Abstractive	Techniques for abstractive summarization are discussed to explain less work has done on abstractive summarizers of Indian languages.	Very few works are carried out in the field of abstractive summarization and there is high need for having research works being carried out in this area.
Bijal Dalwadi et al. 2017 [8]	Marathi	Extractive and abstractive techniques	This survey done the performance analysis of automatic text summarizers for Indian languages it concludes most of the work is done using rule based	The main challenge to summarize the content from several textual and semi structured sources such as databases and web pages,
Kishore Kumar Mamidala et al. 2021 [9]	Telugu, Hindi, Tamil	Extractive and abstractive	This paper provides a survey on text summarization techniques developed for Indian languages	Unavailability of resources like data sets, stop word lists, synsets for the Indian language like Telugu, Hindi, Tamil, etc.
Deepali Kadam et. al. 2015 [10]	Hindi	Extractive, Genetic algorithm, Neural Network, weight learning method	Higher ranked sentences are selected for summary	Sentence scores are calculated using features of each sentences.
Dawinder Kaur et al.	Hindi	Extraction	System removes 30-40% of text to obtain	The summary text obtained by the system

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

2014 [11]			the summary	can be also reducing more to 50% by increasing the minimum weight of the lines
Dhanya P.M. et al. 2013 [12]	Tamil,Odia, Bengali, Punjabi, Kannada, Gujarati	Extractive, Tf-Idf, sentence scoring, graph-based sentence weights	It is proved that methods are not language specific. Not all the features are given equal weights	Same set of sentences in English are used for comparing all the methods
Divakar Yadav et al. 2015 [13]	Hindi	Extractive based on a thematic approach.	System given accuracy is 85%	Evaluation of summary by Human generated summary Language expertise is the requirement
Nikita Desai et al 2016 [14]	Hindi	Extractive approach based on feature vector	The average result of experiments indicates 72% accuracy at 50% compression ratio and 60% accuracy at 25% compression ratio.	The generated summary evaluated against human generated summaries
Malvi shah et al.2019 [15]	Gujarathi	Extractive Graph based ranking model	Using stemmer and String similarity measure summary with good recall can be achieved	Human expertise is the requirement
Meetkumar Patel, 2018 [16]	Gujarathi	Abstractive , Machine learning using neural network	Study explored the machine learning approach to text summarization with neural network	adding diverse sentence simplification techniques is required
Nedunchelian Ramanujam et al. 2016 [17]	Gujarathi	Abstractive and extractive methods	This system gives better outputs than the MEAD algorithm. Results are good for precision, recall, than the existing clustering and lexical chaining method.	System requires knowledge base
Vishal Gupta et al. 2010 [18]	Punjabi	Extractive, feature weight algorithm	Accuracy varying from 81% to 92 0/0.	Domain dependent
Arti Jain et al. 2021[19]	Punjabi	Extractive approach using neural network	The precision, recall and F-measure are achieved as 90.0%, 89.28% an 89.65%	Limited features are required to select for with more understand ability

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

Arti Jain et al. 2021 [20]	Punjabi	Abstractive summarization, Particle swarm algorithm	ROUGE-1 gives better results in comparison to ROUGE-2 Dataset_II having precision as 0.7836, recall as 0.7957 and F-measure as 0.7896 respectively.	Domain dependent
Vishal Gupta et al. 2019 [21]	Punjabi	Hybrid approach based on concept based, statistical, location-based, numeric feature and linguistic features	Statistical features in addition with linguistic, location-based features have improved the precision, recall, ROUGE-2 scores. Overall average F score, Precision, Recall and ROUGE-2 scores increased	Limited dataset of 150 documents
Geetha J.K. et al 2015 [22]	Kannada	Extractive with Singular Value Decomposition (SVD), LSA	LSA used to find similarity between the texts. The accuracy of 94% and precision of 80%	Human expertise required
Arpitha Swamy et al.2019 [23]	Kannada	Extractive	better performance in producing extractive summaries as compared to human summaries	more statistical and linguistic features need to be considered in the process of sentence scoring and sentence ranking.
Sarkar et al. 2012 [24]	Bengali	Extractive	Sentences are ranked based on thematic term and position features.	Need to explore the number of features and apply learning algorithm
Md Ashraful Islam Talukder et al. 2019 [25]	Bengali	Abstractive, encoding and decoding with LSTM.	reduced the train loss to 0.008 and able to generate a fluent short summary	Limited dataset

2.1 Challenges in the Summarization Task

After literature study it is noticed that there are many problems researchers have to face for solving the NLP problems like text summarization. Some major challenges are enlisted below.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

2.1.1 Lack of Standardization

Marathi language uses Devnagari scripts for writing text but single word has different representation in their spellings or aksharas. It is mainly depends on how the language is used by the speaker with its actual intention and tone of speaking like whether it takes long pronunciation (दीर्घ उच्चार) or short pronunciation (ह्रस्व उच्चार) such as for mice (उंदिर) with short pronunciation or mice with long pronunciation (उंदीर), book (पुस्तक) or book (पूस्तक) variation is seen in the written script.

2.1.2 Phrases

A collection of useful phrases in Marathi makes it different kind of language, special and rich to understand too. Because many times a phrase uses some words in the sentence they have special appearances and also their intentions are different than the actual meaning of the language words

Example लेकी बोले सुने लागे

In this, speaker want to target her daughter in laws but she indirectly target to her daughter.

Example हातातील कंगनाला आरसा कशाला?

The above statement means there is no need to understand the situation because everything is very clear to understand.

Example एक ना धड भराभर चिंध्या.

In this, statement meaning is there are many things available in collection but no one is useful.

2.1.3 Post Positions

There are many language words they came with different form having some suffixes added with the basic words.

2.1.4 No Capitalization

In the Marathi language there has no concept of capitalization for writing nouns by default grammar specifications and rules. It create the problem to identify the important words.

2.1.5 Complex Morphology

Due to the complex structures and syntax variations being a free-ordered language Marathi language has grammatical complexities.

2.1.6 Ambiguity

Many times same word express different meaning for that instance and creates complexity to understand the language.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

2.1.7 Code Mixed Data

With original source language other language words or foreign words are used for communication which creates complexity for processing the standard language text. This code mix data is the additional foreign language words rather than Marathi. Ex. Marathi and Hindi, Marathi and English, or Marathi and Konkani etc.

2.1.8 Fast Evolution Rate

Due to everyday technology growth new technical words are continuously added in the original language and treated as an important and integral part of language such as Computer (कॉम्पुटर), Mobile (मोबाईल), Robot रोबोट etc. This is caused to create complexities in language processing tasks.

2.1.9 Language Dialects

Same language is used by different group of people or regions. A region specific language spoken with variations in their vocabulary, tone or sound and compositional structure that also affect the written text. Marathi has many local variations depending on the region and after every 100 kilometres of distance the language dialects are observed. For the word crowd Marathi has various dialects like गर्दा or गर्दी in this, words has different tone and written structure format.

2.2.0 Low Language Resources

For analysis and classification of the language and computational processing sufficient tools are not available. Development of resources are in their initial stages for Marathi like no standard dataset for Marathi, No standard language set for stop words identification, Non efficient language translators, parsers, morphological analyzers etc.

3. DATASET FOR THE PROPOSED SYSTEM

For this proposed system or developmental task we have used own dataset collection that we have obtained online for the experiment. We have used Marathi text stories. These are moral children stories which are downloaded online and processed to accomplish the task. We use this dataset on the trial basis to achieve the summarizer task. This dataset consists of total 1565 number of words. This is created and maintained in the form of text file using Notepad editor and then it is processed using Python and NLTK libraries. Below Table2 shows the sample example and detailed information about the dataset.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

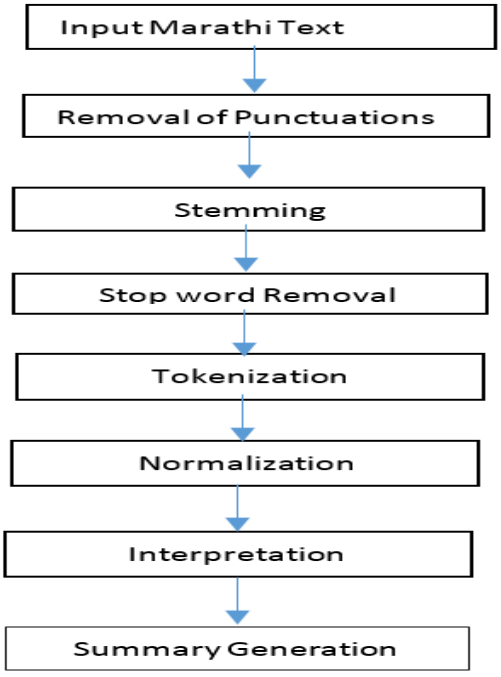
Table 2: Dataset Details for the system

Story Name	Story Label	Story Length in no. of words
उंदराची टोपी	Story1	344
टोपीवाला आणि माकडे	Story2	139
कुत्र्याची हुशारी	Story3	240
धक्का लागल्याने कळून येतं व्यक्तिमत्त्व	Story4	133
गणपतीने आपल्या बुद्धिमत्तेने शर्यत जिंकली	Story5	124

4. PROPOSED METHODOLOGY

The proposed system is implemented using the Python language which is an interpreter high-level, general-purpose programming language. Python has a large set of standard library specially designed for NLP task such as NLTK. We also use the Notepad text editors and SQL server Management studio as the basic requirements. Below Fig.1 shows the main steps and an architectural view of the developmental tasks for the Marathi Summarizer.

Fig 1: Architectural View of the Proposed Summarizer for Marathi



4.1 Input Marathi Text

A Marathi language text is accepted as an input for the proposed system. This is actually a single document source file maintained using Notepad editor and consisting of Devnagari script language words. Below Table3. Shows sample text used for summary generation.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

Table 3: Sample Input Text used for Summary Generation

Title of the story: शिष्याच्या अयोग्य दृष्टीकोनाला गुरूंनी दिलेला सकारात्मक दृष्टीकोन
Original Text एकदा दोन संन्यासी म्हणजेच एक गुरू व शिष्य आठ महिन्यांनी पावसाळ्यात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे. शिष्य म्हणाला, बघा, आपण भगवंताची आठवण करून-करून मरतो त्याचे हे फळ.म्हणून मी सांगतो, प्रार्थना, पूजा इत्यादींत काही अर्थ नाही. दुष्टांचे बंगले चांगले राहिले आणि आपल्या झोपड्या पडल्या. ज्या वेळी तो हे रागाने सांगत होता, त्या वेळी गुरु आनंदाने परमात्म्याला सांगत होते. परमात्म्या, तुझी कृपा म्हणून अर्धे छप्पर अजूनही आहे. शिष्य चिंतेत होता म्हणून त्याला रात्रभर झोप लागली नाही. गुरु पहाटे उठले त्यावेळी अर्ध्या छपरातून चंद्र दिसत होता. हे पाहून ते आनंदी झाले त्यांनी यावर एक कविता लिहिली. त्या कवितेत त्यांनी लिहिले होते, देवा, आम्हाला आधीच ज्ञात असते तर आम्ही आधी अर्धेच छप्पर बांधले असते. आता आम्ही झोपतांना चंद्रही बघू शकतो.

4.2 Pre-processing

It is an important step in text summarization or data mining because we cannot work with raw data. Data pre-processing operation converts raw data into a standard and uniform text format It is necessary to have quality data. It should be checked before applying machine learning or data mining algorithms. In the pre-processing phase different statistical and linguistic features are extracted using special libraries and functions of Natural language tool kit (NLTK) and by implementing the python functions.

4.2.1 Removal of punctuation marks and special characters

A punctuation mark is a symbol or language marker such as a comma (,), Semicolon (;), question mark (?), an exclamatory mark (!), dash (-), hyphen (_), colon (:), apostrophe or single quotation (‘), double quotation (“), dot (.) or a full stop, etc. are specifically used to denote some language impact to create more expressive communication. Below Table4 shows sample Text for removal of punctuation

Table 4: Removal of Punctuation

Input Text	दोन संन्यासी म्हणजे गुरु-शिष्य आठ महिन्यांनी पावसाळ्यात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे.
Output text	दोन संन्यासी म्हणजे गुरु-शिष्य आठ महिन्यांनी पावसाळ्यात परत त्यांच्या झोपड्यांकडे आले आल्यावर त्यांनी बघितले की झोपड्यांचे अर्धे छप्पर उडून गेले आहे

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

4.2.2 Stop Word Removal

In this, frequently occurred words are removed from the text. Below Table5 shows Stop word removal.

Table 5: Removal of Stop words

Input Text	दोन संन्यासी म्हणजे गुरु शिष्य आठ महीन्यांनी पावसाळ्यात परत त्यांच्या झोपड्यांकडे आले आल्यावर त्यांनी बघितले की , झोपड्यांचे अर्धे छप्पर उडून गेले आहे
Output text	संन्यासी गुरु शिष्य आठ महिन्यांनी पावसाळ्यात परत झोपड्यांकडे आल्यावर बघितले झोपड्यांचे अर्धे छप्पर उडून

4.2.3 Tokenization

It is the process of splitting the whole text into number of individual sentences or number of individual words. It is important to understand the actual structure of the word, its meaning, and its position wise impact on the whole language content and also to process the language content properly after its exact analysis.

4.2.3.1 Sentence Tokenization

It is also called sentence segmentation. In this, whole the paragraph or source text is splited into number of individual sentences. Each the sentence from the source input is separated using a comma and is quoted with the single apostrophe symbol. Below Table6 shows sentence tokenization

Table 6: Sample Example for Sentence Tokenization

Input Marathi Text	एकदा दोन संन्यासी गुरु व शिष्य आठ महिन्यांनी पावसाळ्यात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बेघितले की , झोपड्यांचे अर्धे छपर उडून गेले आहे.
Output Sentence Tokenization	‘एकदा दोन संन्यासी गुरु व शिष्य आठ महिन्यांनी पावसाळ्यात त्यांच्या झोपड्यांकडे आले.’,’ आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे.’

4.2.3.2 Word Tokenization

In this, each individual sentence is divided into number of words within it. Each this word is also separated using a comma and enclosed within a single quotation mark. Each this word is a separate token used for the further computational processing. This operation can be done using NLTK functions. Below Table7 shows Word tokenization.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

Table 7: Sample Example for Word Tokenization

Input Marathi Text	एकदा दोन संन्यासी गुरु शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले. आल्यावर त्यांनी बघितले की , झोपड्यांचे अर्धे छपर उडून गेले आहे.
Output word Tokenization	'एकदा', 'दोन',' संन्यासी','गुरु',' शिष्य', 'आठ ',' महिन्यांनी','पावसाळयात', 'परत', 'त्यांच्या',' झोपड्यांकडे','आले.','आल्यावर',' त्यांनी',' बघितले',' की' ,' झोपड्यांचे',' अर्धे','छपर', ' उडून',' गेले',' आहे.'

4.2.4 Stemming

In this basic operation a word with added suffixes are identified and then by removing the suffixes words are formed with their basic root form. This is done using a suffix removal algorithm or stemmer for Marathi.

Table 8: Sample Example for Stemming

Input Marathi Text	एकदा दोन सन्यासी गुरु शिष्य आठ महिन्यांनी पावसाळयात परत त्यांच्या झोपड्यांकडे आले . आल्यावर त्यांनी बघितले की, झोपड्यांचे अर्धे छप्पर उडून गेले आहे.
Output stemming	एकदा दोन संन्यासी गुरु शिष्य महिना पावसाळयात परत त्यांच्या झोपड्या आले. आल्या त्यांनी बघितले की झोपडी अर्धे छप्पर उडून गेले आहे

4.2.5 Normalization

In this phase normalized text is obtained after the preprocessing and with some manual validations. There are many variations in Marathi for the same word. By following the manual validation text, it is verified for the processing. The text is validated for pure Marathi language words for the computational processing. Impurities are removed from the original language.

4.3 Processing

In this, we have used machine Learning algorithms, mathematical modelling, and statistical knowledge, linguistic techniques together so that entire process can be automated. The processing phase is applied to extract various statistical and linguistic features.

4.4 Interpretation

In this phase, data interpretation is done. It is important to assign meaning to the collected information and determining the conclusions. Implications of the findings. It is useful for qualitative and quantitative analysis of the data. Abstract summaries need to go through the interpretation phase in these different contents are combined to form a general content or meaning.

ISSN 1533-9211

DOI 10.5281/zenodo.7073510

4.5 Summary Generation

Finally, the summary is generated for the given Input document using abstract method. In this, system generate natural language text based on previous steps. It is also in the form of Marathi text obtained in shorter version of the document. Below Table 2. Shows a sample result for a Marathi story document.

5. RESULTS AND DISCUSSION

Our system generated summary has ability to produce the summary in shorter version. System generated summary is evaluated with the Human generated summary. With the help of manual similarity measures of comparison of human generated summary and system generated summary. Our system gives 70% correct summary. Our system is capable of reducing 44.38 % of the text. For evaluation of the summary, we have used formula for accuracy. It is found that system gives a better response to generate the abstract summary. Below table3 shows the sample result generated by our Summarizer.

$$(\text{Accuracy}) = \frac{\text{total No.of correct words}}{\text{total No. of correct words + tatal no.of incorrect words}} * 100$$

Label of the story	Length of source document in No. of Words	Length of Summary Text in No. of Words	Percentage of reduction in text	Average Percentage Count for reduction in text by the system
Story1	344	186	54.06	44.48
Story2	139	52	37.41	
Story3	240	100	41.66	
Story4	133	48	36.09	
Story5	124	60	48.38	
Story6	142	74	52.11	
Story7	145	70	48.27	
Story8	150	56	37.33	
Story9	130	63	48.46	
Story10	118	52	44.06	

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

CONCLUSION AND FUTURE WORK

For some Indian languages like Bengali, Malayalam, Hindi, Odia, Gujarati, Tamil, Telugu, Gujarati and Punjabi etc. good work has been done but for Marathi less work has been done. For Marathi, it is in its initial stage of development for the various NLP applications like Summarizer. It is noticed that different combination of features and techniques caused to works differently for different types of content.

Hence, it is challenging to create a single document text summarizer for the Marathi language. Our system generates abstract summary which can reduce 44. 48% of text from original text as a summary. In future, we are aiming to use more features and enhancement of the techniques for abstracting Marathi text. As well as enhancing the database for to improve the accuracy of the summary text. Also, we will try different machine learning techniques for summarization to achieve better accurate results. There are several text summarization algorithms were proposed for the automatic summarization of documents. We are trying to develop a system which is comparatively more capable and efficient for summarizing Marathi text. Our system gives accuracy 44.48 % to reduce the original text contents.

Acknowledgements

Authors would like to acknowledge and thanks to CSRI DST Major Project sanctioned No.SR/CSRI/71/2015 (G), Computational and Psycholinguistic Research Lab Facility supporting to this work and Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India. Also thankful to Chhatrapati Shahu Maharaj Research Training and Human Development Institute (SARTHI), Pune for providing financial assistance for this Ph. D. research work. I would like to express my sincere thanks to research guide Dr. C. Namrata Mahender (Asst. Professor) of the Computer Science and IT Department, Dr. B. A. M. U., Aurangabad, for providing research facilities, constant technical and moral support.

References:

Namisha Dheer, Chetan Kumar. Automatic Text Summarization: A detailed study. International Journal of Science and Research (IJSR), https://www.ijsr.net/get_abstract.php?paper_id=NOV161195, Volume 5 Issue 2, February 2016, 429 - 433, #ijsrnet

Gaikwad, D.K., Mahender, C.N.: A review paper on text summarization. Int. J. Adv. Res. Comput. Commun. Eng. 5(3), 2319–5940 (2016). ISSN (Online) 2278–1021. 2016.

Archana AB, Sunitha C. An overview on Document Summarization Techniques. International Journal on Advanced Computer Theory and Engineering (IJACTE). 2013

Jagadish S. Kallimani, Srinivasa K. G., Eswara Reddy B.. Information Extraction by an Abstractive Text Summarization for an Indian Regional Language. IEEE 2011.

Dhanya P. M., Jethavedan M. Comparative Study of Text Summarization in Indian Languages. International Journal of Computer Applications (0975 8887) Volume 75 No.6, August 2013.

Nikita Munot, Sharvari S. Govilkar. Comparative Study of Text Summarization Methods. International Journal of Computer Applications (0975 8887) Volume 102 No.12, September 2014.

Sunitha, A. Jaya, Amal Ganesh. A Study on Abstractive Summarization techniques in Indian Languages. C. Sunitha et al. / Procedia Computer Science 87 (2016) 25 — 31.doi:10.1016/j.procs.2016.05.121

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

Bijal Dalwadi, Nikita Pate1, Sanket Suthar. A Review Paper on Text Summarization for Indian Languages. IJSRD - International Journal for Scientific Research & Development Vol. 5, Issue 07, 2017 | ISSN (online): 2321-0613.

Kishore Kumar Mamidala. Text Summarization for Indian Languages: A survey. International Journal of Advanced Research in Engineering and Technology (IJARET) Volume 12, Issue 1, January 2021, pp. 530-538, Article ID: IJARET 12 01 049.

Deepali P Kadam, Nita Patil, Archana Gulathi. A Comparative Study Of Hindi Text Summarization Techniques. Genetic Algorithm and Neural Network, International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 - 8616 Volume 4, Special Issue March 2015.

Dawinder Kaur, Rajbhupinder Kaur. Automatic Summarization of Text Documents Written in Hindi Language. International Journal of Computer Science and Mobile Computing, ISSN 2320-088X IJCSMC, vol. 3, Issue. 10, pg.320 - 323, October 2014.

Dhanya P. M., Jethavedan M. Comparative Study of Text Summarization in Indian Languages. International Journal of Computer Applications (0975 8887) Volume 75 No.6, August 2013.

Divakar Yadav, Vimal Kumar K. An Improvised Extractive Approach to Hindi Text Summarization. Second International Conference on Information Systems Design and Intelligent Applications. 2015 http://dx.doi.org/10.1007/978-81-322-2250-7_28

Nikita Desai, Prachi Shah. Automatic text summarization using supervised machine learning technique for hindi language IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.2016

Malvi Shah, Dr.Kalyani Patel, Gujarati Text Summarizer International Research Journal of Engineering and Technology (IRJET) e-SSN: 2395-0056 Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072.2019

Meetkumar Patel, Adwaita Chokshi, Satyadev Vyas, Khushbu Maurya. Machine Learning Approach for Automatic Text Summarization Using Neural Networks. IJARCCE ISSN (Online) 2278-1021 ISSN (Print) 2319-5940 International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 7, Issue 1, January 2018.

Nedunchelian Ramanujam and Manivannan Kaliappan. An Automatic Multidocument Text Summarization Approach Based on Naïve Bayesian Classifier Using Timestamp Strategy.Research Article | Open Access Research Article | Open Access Volume 2016 |Article ID 1784827 | https://doi.org/10.1155/2016/1784827.2016

Vishal Gupta, Gurpreet Lehal. A Survey Of Text Summarization Extractive Techniques. Journal Of Emerging Technologies In Web Intelligence · August 2010.

Arti Jain, Anuja Arora , Divakar Yadav , Jorge Morato , and Amanpreet Kaur. Text Summarization Technique for Punjabi Language Using Neural Networks. The International Arab Journal of Information Technology, Vol. 18, No. 6, November 2021.

Arti Jain, Divakar Yadav, Anuja Arora. Partical Sworm Optimization. https://www.igi-global.com/gateway/article/full-text-html/275001&riu=true, 2021

Vishal Gupta, Narvinder Kaur(2019). A Novel Hybrid Text Summarization System for Punjabi Text. DOI 10.1007/s12559-015-9359-

Geetha J. K., Deepamala N. Kannada Text Summarization using Latent Semantic Analysis https://sci-hub.hkvisa.net/10.1109/ICACCI.2015.7275826

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7073510

Arpitha Swamy, Srinath S. Automated Kannada Text Summarization using Sentence Features. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019

Sarkar K., Kamal. An approach to summarizing Bengali news documents. In proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 857-862. ACM, 2012.

Md Ashraful Islam Talukder, Sheikh Abujar , Abu Kaisar Mohammad Masum Bengali abstractive text summarization using sequence to sequence RNNs. 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). At: Kanpur, India, 2019.