Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

Marathi e-Newspaper Text Summarization Using Automatic Keyword Extraction Technique

[1]Mr. Shubham Bhosale, [2]Ms. Diksha Joshi, [3]Ms. Vrushali Bhise, [4]Prof.Rushali A. Deshmukh

1,2,3,4 RajarshiShahu College Of Engineering, Pune

Abstract:- *In today's world, Time is of essence and no one wants to waste their time not even on knowing what is happening around him. most used and reliable way to forward news is newspaper until television was invented. no one wants to read a half page article describing whole event of any incident happened around them. Everyone wants that everything around them should be as fast as possible. So how can we make newspaper articles as small as possible and easy to read and understand. Summarization is the process of minimizing the text content from a larger text document and displaying only the information which is most important. The intention of text summarization is to express the content of a document in a condensed form that meets the needs of the user. Most extractive summarization techniques revolve around the concept of finding keywords and extracting sentences that have more keywords than the rest. Keyword extraction usually is done by extracting relevant words having a higher frequency than others. Manual extraction or annotation of keywords is a tedious process brimming with errors involving lots of manual effort and time. In this paper, we proposed an algorithm to extract keyword automatically for text summarization in e-newspaper articles.*

Index Terms*:-Automatic keyword detection, e-Newspaper, Natural language processing, Text summarization.*

Introduction

With the ever-growing popularity of an online newspapers is accepted by user. There are many popular Marathi e-newspapers available freely in the internet, such as Maharashtra Times, Lokmat, Sakal, Prabhat, Pudhari etc. These newspapers are extracting all the necessary information from newspapers. Extracting information from newspapers is a difficult job for every person. There is a need for a tool that extracts only needed information from these data sources. Automatic keyword extraction is the process of selecting words and phrases from an article that can at best project the core sentiment of the article without any human intervention depending on the model . Summarization is a process where the most salient features of a text are extracted and compiled into a short abstract of the original wording.
There are mainly two types of summarization

1.Extractive summarization[4]
2.Abstractive summarization.[4]

In Extractive summarization, important sentences are identified and only those sentences are included in summary, desired summary length is obtained by use of compression ratio. In case of abstractive summarization, according to scoring criteria recognize relevant sentences and process these sentences so as the sentences can be included in summary. Abstractive summarization includes deep understanding of natural language and it is also based on compression. Mostly automatic summarization deals with extractive type of summarization.

The system works by assigning scores to sentences in the document to be summarized, and using the highest scoring sentences in the summary. Score values are based on features extracted from the sentence. A linear combination of feature scores is used. The intended user is considered to have little background knowledge or reading ability. The system helps by simplifying the individual words used in the summary and by drawing the pre-requisite background information from the web.

Literature survey

In this section we are discussing different approaches for Keyword extraction, Keyword ranking/classification ,Text summarization and algorithms to do so.

Paper no. 1

In [1], authors Reddy Naidu1, Santosh Kumar Bharti1, KorraSathya Babu1, and Ramesh
Kumar Mohapatra propose a technology which can work on telugunews papers and summarize their content. In their approach they need human intervention to train the system to find the probable keywords. These keywords are then passed to the POS tagger to further analyse.in learning stage system uses newspaper cut-outs, this cut-out is considered as target document and useful statistics such as nouns, verbs, adverbs, etc. are calculated. By using keyword extraction

algorithm key phrases are extracted by using probability distribution, and by using those key phrases new summarized document.

Paper no. 2

In [2], authors M. Hanumanthappa, M Narayana Swamyand N M Jyothi explained that
For doing effective text summarization, before doing extraction text preprocessing is necessary means the data which is in unstructured form should be first converted to structured form. They further discuss about the four stages which are needed for txt preprocessing, which are selecting candidate terms, filtering, vector space model, and ranking. For selecting candidate terms text is tokenized and candidate terms are selected by using n-gram approach. In filtering vocabulary pruning is done to further minimize the candidate terms and predefined common terms are removed from the list. In vector space model the text document is represented in a form of vector. The main reason to use VSM is to identify the actual meaning of an term according to the context. And ranking deals with the statistics creation process.

Paper no. 3

In [3], authors Gabriel Silva, Rafael Ferreira, Rafael Lins, Luciano Cabral, Hilário Oliveira,
Steven J. Simske, Marcelo Riss discussed about an grammatically correct English summery creation system which was developed by Lins. This system works on the news text extracted from the CNN website. The system mainly produces 3-5 lines news highlights rather than producing a whole summarized article. This is done by adding all news data into an XML document and numbering each sentence and paragraph. Tis XML document is then passed to the feature extractor where a feature vector is generated by using it. And this vector is used for calculating standards such as aggregate similarity , sentence length, proper nouns, TF/IDF , Uppercases, Word Frequencies. And by using these standards 3-5 sentences are formed and displayed to the user.

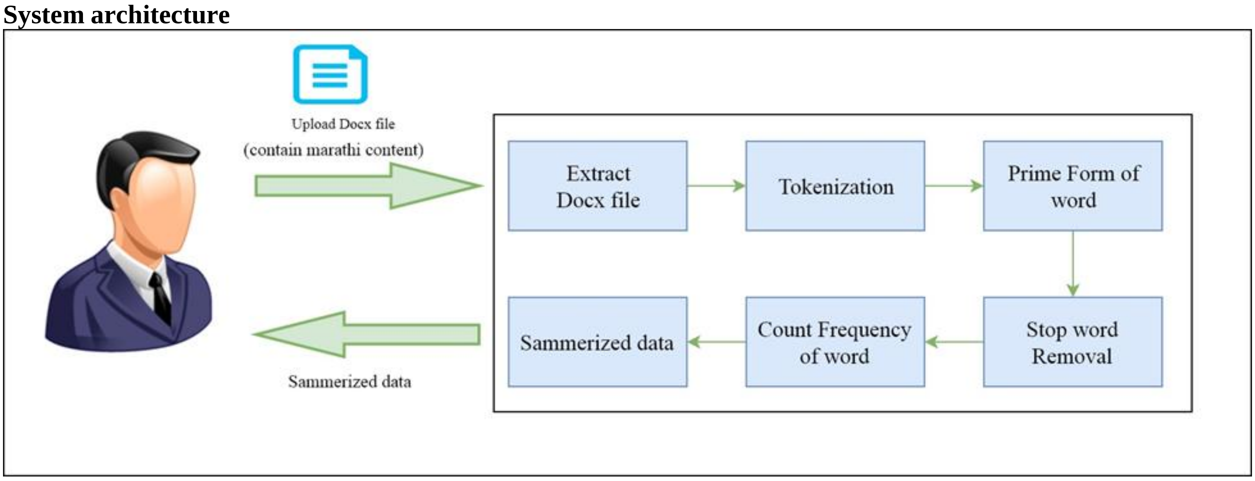
Paper no. 4

In [4], authors Sheetal Shimpikar, Sharvari Govilkar discuss about technologies using which we can develop text summarizer for Indian regional languages. In this paper they discuss about two main summarization techniques which are Abstractive and Extractive summarization and explain in detail how they work. Abstractive text summarization contains different methods such as Structure based approach, Tree based method, Template based method, Ontology based method, Rule based method, Semantic based approach. And extractive text summarization techniques such as Term frequency Inverse Document Frequency, Cluster based method such as k-means algorithm, Maximum , relevance multi document(MMR-MD), and graph theoretic. And each technology works differently depending upon the content provided to it.

Proposed system

In this section we will know about how the task of keyword extraction, summarization is done in the system. Tasks are further divided for simplification of development process. Starting with accepting the article the user wants to summarize, it can be done by accessing article directly from newspaper site, by uploading the doc file, or by copy pasting it onto the site. After accepting data, it is analysed for finding frequently occurring words and classifying them. Using these classified words we pick the lines containing those words to create a summarized article, which is main objective of this system.

System architecture

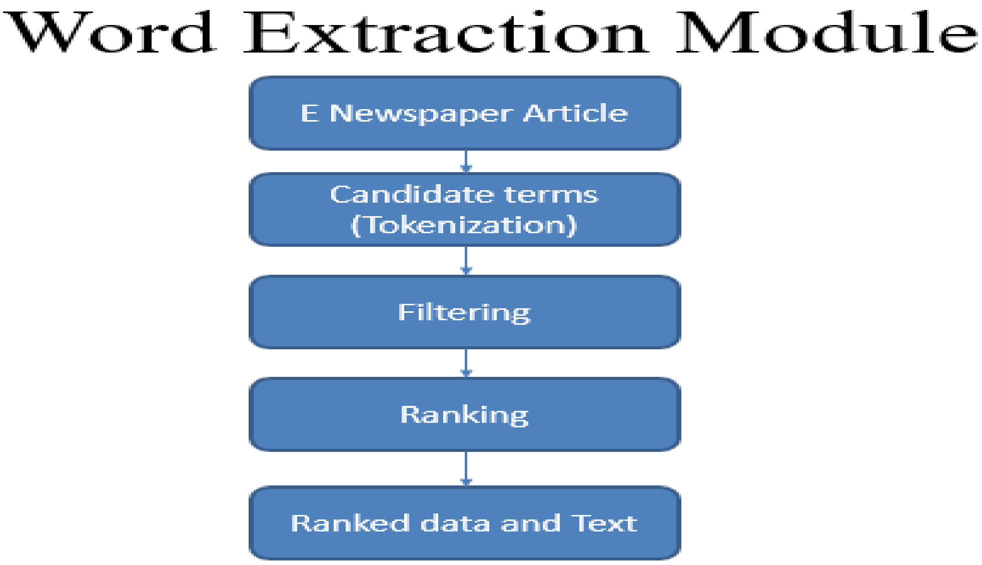


Modules

System is made up of two main modules.

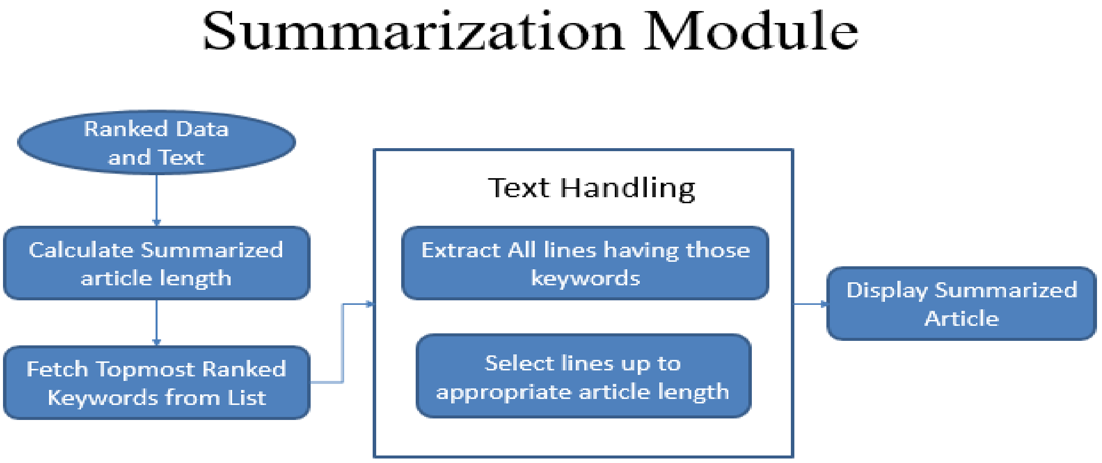
Word Extraction Module[6]

This module mainly deals with text provided by the user. First of all text provided by the user is tokenized for ease of further word processing. After getting tokenized data it is filtered to remove punctuations and numbers. Filtered data is then ranked and components such as stopword list, frequencies, etc. are created.



Summarization Module

As name suggests this module deals with text summarization. First it calculates average article length which can be calculated and it is average of 30% to 40 % size of original article. Then according to size calculated in previous step topmost ranked keywords are selected which can be in range of 5-15 keywords. Then all the lines containing those keywords are fetched and a new summarized article is displayed to the user.


Summarization Module
Ranked Data and Text
Calculate Summarized article length
Fetch Topmost Ranked Keywords from List
Text Handling
Extract All lines having those keywords
Select lines up to appropriate article length
Display Summarized Article

Algorithms

Algorithm 1: Extract_Keywords[7]

```
Data: doc := Input Article
P(Tag) := Set of Trained Probabilities
Num_Keywords := Required Number of
Keywords
Result: Keywords[]
Res_doc:=resolve(doc)
Pos_doc:=pos_tagger(res doc)
top:=0
while word in pos_doc do
f lag:=0
for i ← 0 to top do
if word.text=wordset[i].text and
word.tag=wordset[i].tag then
```

```
wordset[i].count:=wordset[i].count+1
f lag:=1
end
end
if f lag=0 then
wordset[top + 1].word:=word.word
wordset[top + 1].tag:=word.tag
wordset[top + 1].count:=1
wordset[top + 1].score:=0 top:=top+1
end
end
for i ← 0 to size do
wordset[i].score:=wordset[i].count*P(wordset[i].tag)
end
sort_desc(wordset.score)
for i ← 0 to Num_Keywords do
Keywords[i]:=wordset[i]
End
```

Algorithm 2: Stop Words Removal Approach[7]

Step 1: The target document text is tokenized and individual words are stored in array.

Step 2: A single stop word is read from stopword list.

Step 3: The stop word is matched with the target text which is in form of array by using sequential search technique.

Step 4: If it matches, the word in array is removed, and the comparison is continued till length of array.

Step 5: After removing one stopword completely, next stopword is taken from stopword list and again Process follows step 2. The algorithm runs until all the stopwords are matched with text and removed.

Step 6: Resultant text after operation is displayed, also required statistics like stopword removed, no. of stopwords removed from target text, total count of words in target text, count of words in resultant text, individual stop word count found in target text is displayed.

Conclusion

In this paper, the proposed work deals with e newspaper articles for summarization. The keyword extraction algorithm works to find the top scored words efficiently, and by using this data the summarization module produces summarized article which mainly depend on the size of original article. For selecting words statistical approach is used due to its better performance and less complexity.

References

1. Reddy Naidu1, Santosh Kumar Bharti1, KorraSathya Babu1, and Ramesh Kumar Mohapatra,” Text Summarization with Automatic Keyword Extraction in Telugu e-Newspapers”. National Institute of Technology, Rourkela, Odisha, India (2014)

2. M. Hanumanthappa, M Narayana Swamyand N M Jyothi”Automatic Keyword Extraction from Dravidian Language” ISSN 2348 – 7968(2014)

3. Gabriel Silva, Rafael Ferreira, Rafael Lins, Luciano Cabral, Hilário Oliveira, Steven J. Simske, Marcelo Riss”Automatic Text Document Summarization Based on Machine Learning”

4. SheetalShimpikar, SharvariGovilkar, A Survey of Text Summarization Techniques for Indian Regional Languages , International Journal of Computer Applications (0975 –8887) Volume 165 –No.11, May 2017

5. Litvak M. and Last M.: Graph-based keyword extraction for single-document summarization. In Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pp. 17{24, ACL (2008)

6. Michael J. Giarlo. A comparative analysis of keyword extraction techniques. Rutgers,The State University of New Jersey and Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, Bo Wang.

7. Y. Matsuo, M. Ishizuka. Keyword extraction from a single document using word co-ocuurrence statistical information. International Journal on Artificial Intelligence Tools, 2004]