Project Report on

Text Summarization Using NLP

For course Natural Language Processing

By

KHUSHBOO LODAYA	BE-3	20
DHARMI MEHTA	BE-3	23
NINAD MANKAR	BE-3	22

Guide

Ms. Dipti Mukadam

DEPARTMENT OF COMPUTER ENGINEERING

SHAH AND ANCHOR KUTCHHI ENGINNERING COLLEGE

CHEMBUR, MUMBAI – 400088.

2019 – 2020

ABSTRACT

In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently . It is very difficult for human beings to manually extract the summary of large documents of text. There are plenty of text materials available on the internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it. In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings. In our approach, we have used an extractive summarization method consists of selecting important sentences from the original document and concatenating them into shorter form. In our proposed system, we are using numerical methods to extract summary which has an advantage that we don't need any previous data for summary extraction. The importance of sentences is decided based on score of numerical features of sentences like TF-ISF, Sentence Length, Sentence Position, Sentence Similarity, numerical data.

Keywords – Text Summarization, Knowledge bases, Sentence Similarity, Hindi Summarization, extractive summarization.

TABLE OF CONTENT

1. Title Page 1
2. Abstract 2
3. Table of Contents 3
4. Introduction 4
5. Literature Survey 5
    - Survey of Existing System 5
    - Limitations of Existing System 6
    - Problem Statement 6
    - Objective 6
    - Scope 7
6. Proposed System 8
    - Analysis/Framework/Algorithm 8
    - Methodology 9
7. Implementation Details 10
    - Modules and Descriptions 10
    - Snapshots 13
8. Result and Analysis 15
9. Conclusion and Future Scope 16
10. References 17

INTRODUCTION

Before going to the Text summarization, first we, have to know that what a summary is. A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is, it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then expresses those concepts in clear natural language.

With the steady progress in the field of technology, the internet's growth has also increased in a tremendous rate. People can find hoards of information easily in various forms like text document, statistics and data. It is usually known that the internet provides more than required amount of information. In that way more than one problem were recognized searching for important information through a profuse quantity of documents available and absorbing a large amount of relevant information [1]. Previously storage of large data files were challenging and if could replace these large document files with their summaries then we may overcome this downside. To produce a summary of a large text document we need a reader and an identifier to select between unnecessary and prime words/sentences in the text file cluster to generate summary. A summary which states the gist of the document helps in finding relevant information quickly. Document summarization also provides a way to cluster similar document and present a summary [4]. Automatic document summarization is a primary analysis area in natural language processing (NLP). Natural language processing, comes under the field of science, technology and artificial intelligence and machine learning with the interactions between computers and human language. Generation of summaries without the use of NLP may lack semantic and cohesion.

LITERATURE SURVEY

SURVEY OF EXISTING SYSTEM

Many previous works on extractive summarization uses mainly two major steps:(1) ranking the sentences based on the score which are computed by combining few or several features such as term frequency(TF), position information and cue phrases(Baxendale, 1958);(Luhn, 1958) and (2)selecting few top ranked sentences to form summary.

The very first work in automated text summarization was done by (Luhn, 1958). He used word frequency (number of times a word occurs in a document) and phrase frequency as features to produce summaries. It has been assumed that the most frequent words are indicative of the main topic of a document.

Although subsequent research has developed several summarization methods based on the new features, the work presented by (Baxendale, 1958) is still used today as a foundation of extraction based summary. P.B. Baxendale (Baxendale, 1958) proposed novel feature that is sentence location or sentence location position in an input document. It is analyzed that sentences which are positioned at the beginning or at the end of the document are more important than other sentences in the document.

H.P. Edmundson (Edmundson, 1969) proposed a novel structure for a text summarization. They proposed two new features. First, cuewords that is presence of most indicative words into a document such as finally, in summary, lastly, etc. Second, straightforward feature title or heading words for which an additional weight was assigned to a sentence, if sentences have heading words in it.

Later, (Kupiec, Pedersen, & Chen, 1995) proposed a machine learning approach to text summarization. They described a new technique of summarization with naive-Bayes classifier, the classification function classifies each sentence as whether it is extraction worthy or not.

LIMITATION OF EXISTING SYSTEM

There are a number of limitations pertaining to some approaches. Recent studies have attempted to address some of these limitations. The next big challenge is not only to focus on the summary information content, but efforts should also be put into the readability aspect of the generated summary itself. The future trend of automatic text summarization is most likely to move along this direction. Sometimes useful information item is difficult in constructing grammatically correct sentences. It reduces the linguistic quality of summary. Text Summarization requires lot of processing time. Document with similar content but different vocabularies may result in a false negative match. System for handling semantic content may use special tags. It doesn't handle synonymy (same meaning) and polysemy (a word with different or multiple meaning in different context).

PROBLEM STATEMENT

Automatic text summarization is a technique which compresses large text into a shorter text which includes the important information. The computer program is given a text and it returns a summary of the original text. This is done by reducing redundancy of the text and by extracting the essence of the text. In this project we understand text summarization and create our own text summarizer in python for Hindi corpus.

OBJECTIVE

The aim of summarizing extractive documents is to select automatically a number of indicative sentences, passages or paragraphs from the original document and reducing a long paragraph of text with a computer program in order to preserve the description that returns the most relevant points of the original text or document. Hindi Automatic text summarization is exactly meant for the same. It provides the reader with filtered description of source text and a non-redundant presentation of facts found in the text.

SCOPE

Business leaders, analysts, students and academic researchers need to go through huge numbers of documents every day to keep ahead, and a large portion of their time is spent just figuring out what document is relevant and what isn't. By extracting important sentences and creating comprehensive summaries, it's possible to quickly assess whether or not a document is worth reading. It can be used to summarize a technical paper, provide movie review, headlines of news, etc.

PROPOSED SYSTEM

ANALYSIS/ALGORITHM/FRAMEWORK

The proposed technique can be grouped into two major steps named as Preprocessing Step, Processing Step and Extraction Step which are explained below in details.

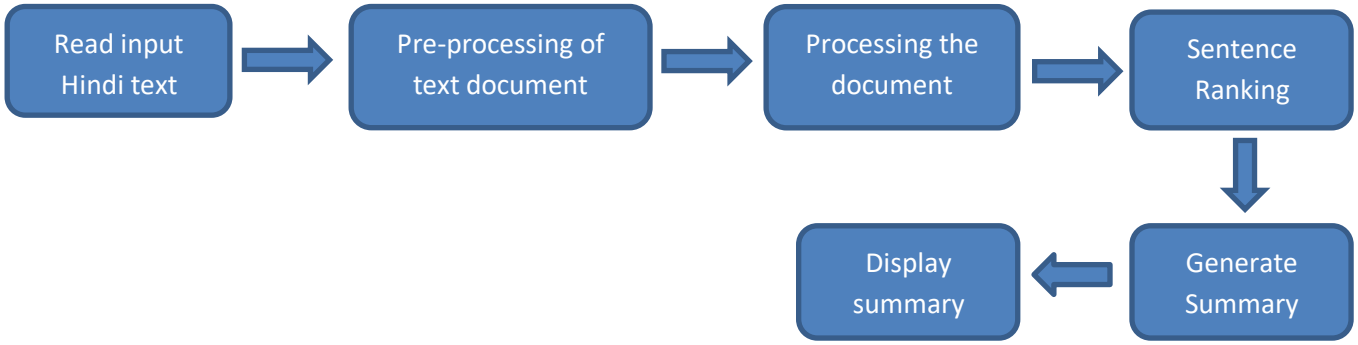

Fig. 1 Flowchart of The Model

METHODOLOGY

1. Read Input text file.
2. Preprocess the file. (Processing step)
   I. Sentence Segmentation.
   II. Tokenization of each sentence to words.
   III. Stop Words Removal.
   IV. Stemming of Words.
3. Processing Step (feature extraction + sentence ranking)
   I. Extract following features from text file
      i. Average TF-ISF
      ii. Sentence length
      iii. Sentence Position
      iv. Sentence Similarity

v. Numerical Data

II. Sentence ranking to rank sentences in range of "0 to 1" with 1 indicating most important and 0 indicating not important sentence based on the each feature's normalized scores.

4. Generate Summary

While (Sentences in Summary file does not exceed maximum limit as per given by compression ratio) extract all sentences from the file sort by sentence rank of Maximum rank (rank5) to minimum (rank1).

IMPLEMENTATION DETAILS

MODULES AND DESCRIPTION

1. **Pre Processing Step**

This intermediate preparation stage is called a Preprocessing step which is a structured representation of the original text. In the preprocessing step the input text obtained from the text file first split into sentences using segmentation method, then sentences are further split into words using tokenization and then stop-words are removed to clean the original text.

i. Sentence Segmentation

It is boundary detection for a sentence. In Hindi, sentence is segmented by identifying boundary of sentence that ends with (|).

ii. Tokenization

In tokenization the sentences are broken up into discrete bits or tokens (words). It omits certain characters, such as punctuation, spaces and special symbols between words.

iii. Stop-Words Removal

Stop-Words include function words, articles, prepositions, conjunctions, prefix, postfix, etc. i.e. common words that carry less important meaning than keywords. So these types of words should be removed from input text document, otherwise the sentence having more no of stop-words could have higher weight.

iv. Stemming

In Stemming process, the suffixes are ignored and removed from words to get the common origin. It recognizes words with common meaning and form as being identical. Syntactically similar words, such as plurals, verbal variations, etc. are considered similar. e.g. “walk”, “walking” and “walked” are counted as same and derived from a stem word “walk”.

2. **Processing Step**

In this we decide and calculate the features that affect the relevance of sentences and then weights are assigned to these features using weight learning method. Higher ranked sentences are extracted for summary.

i. Average TF-ISF

TF-ISF stands for term frequency-inverse Sentence frequency and the TFISF weight is a numerical measure used to evaluate how important a word is to a document.

$$TF(t,S) = \frac{\text{No of times term (t) appears in a Sentence S}}{\text{Total number of Words in the Sentence d}}$$

$$ISF(t,S) = \log\left(\frac{\text{Total No. of Sentences}(|S|)}{\text{Number of Sentences containing the term t}}\right)$$

The score of a sentence k is computed based on the frequency of important words occurrence in a sentence.

$$AvgTFISF(S_t) = \sum TF * ISF$$

ii. Sentence Length

The short sentences such as datelines and author names are not expected to belong to the summary. In the same way, too long sentences may contain a lot of redundant data and hence are unlikely to be included in the summary. So, we eliminate the sentences which are too short or too long.

L = Length of the sentence

MinL = Minimum length of sentence

MaxL = Maximum length of

Minθ= Minimum Angle(0 deg)

Maxθ= Maximum Angle(180 deg)

SL = 0, If (L < MinL) or (L >MaxL)

Otherwise

SL = sin[(maxθ −minθ) (maxL−minL) ×(L−minL)]

iii. Numeric Data(ND)

Usually the numerical data is used to show the important mathematical or statistical analysis providing some vital information in a document and hence claims to be a part of summary with its essential contribution to the document.

$$ND = \frac{Number\ of\ Numeric\ Data\ in\ Sentence}{Sentence\ Length}$$

iv. **Sentence Position**

Usually, sentences in the beginning define the theme of the document, while end sentences conclude or summarize the document. So, position of the sentence in the text, decides its importance. Threshold value in percentage, defines how many sentences in the beginning and at the end are retained in summary with weight SP=1

v. **Sentence to sentence similarity**

For each sentence S compute the similarity between S by creating a weighted matrix with other sentences S' of the document, then add up those similarity values. It gives us the raw value of this feature for S. There are many approaches to calculate the similarity between two sentences.

$$SS = \sum_{1}^{N} Sim(i, j) When, i \neq j$$

$$Sim(i, j) = \text{Number of Words overlapping between sentence S(i) and S(j)}$$

3. **Sentence Ranking**

Sentence ranking is the very important step to determine which sentence should be included in our summary. In our proposed method sentence ranking per sentence is evaluated using individual score obtained from each feature per sentence and then overall collective score of each sentence is computed by adding up individual score of each features. Then each score of sentences is normalized between 0 - 1. Sentences are sorted based on the descending order of score values. Depending on the compression rate, sentences are extracted from the document to generate summary.

4. **Generate Summary**

Using ranking for each sentence obtained from above sentence ranking step, depending upon the user compression rate input all sentences are extracted till sentence in the summary file

does not exceed maximum limit. Then summary is generated in the order of original text document.

SNAPSHOTS

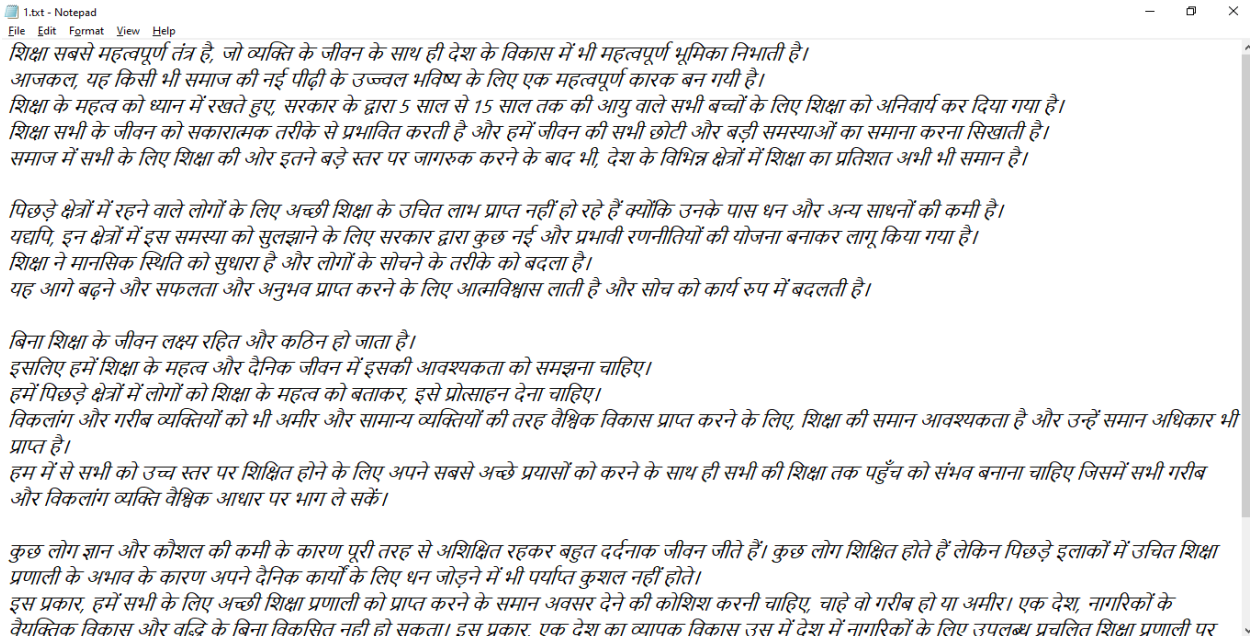

Fig. 2 Original Hindi Data

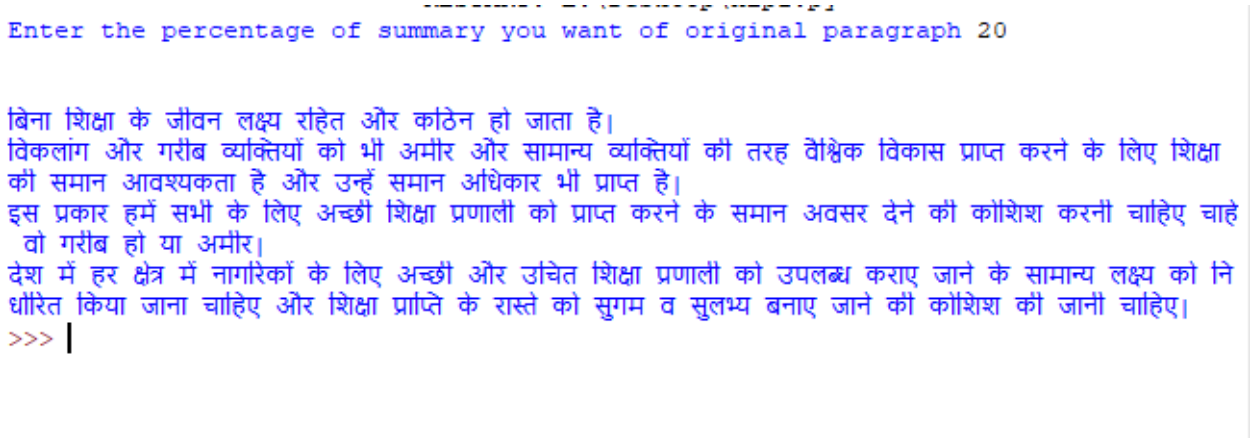

Fig. 3 Paragraph after 20 percent compression ratio.

```
Enter the percentage of summary you want of original paragraph 40


शिक्षा सबसे महत्वपूर्ण तंत्र है जो व्यक्ति के जीवन के साथ ही देश के विकास में भी महत्वपूर्ण भूमिका निभाती है।
बिना शिक्षा के जीवन लक्ष्य रहित और कठिन हो जाता है।
इसलिए हमें शिक्षा के महत्व और दैनिक जीवन में इसकी आवश्यकता को समझना चाहिए।
विकलांग और गरीब व्यक्तियों को भी अमीर और सामान्य व्यक्तियों की तरह वैश्विक विकास प्राप्त करने के लिए शिक्षा
की समान आवश्यकता है और उन्हें समान अधिकार भी प्राप्त है।
 कुछ लोग शिक्षित होते हैं लेकिन पिछड़े इलाकों में उचित शिक्षा प्रणाली के अभाव के कारण अपने दैनिक कार्यों के
लिए धन जोड़ने में भी पर्याप्त कुशल नहीं होते।
इस प्रकार हमें सभी के लिए अच्छी शिक्षा प्रणाली को प्राप्त करने के समान अवसर देने की कोशिश करनी चाहिए चाहे
 वो गरीब हो या अमीर।
 इस प्रकार एक देश का व्यापक विकास उस में देश में नागरिकों के लिए उपलब्ध प्रचलित शिक्षा प्रणाली पर निर्भर
करता है।
देश में हर क्षेत्र में नागरिकों के लिए अच्छी और उचित शिक्षा प्रणाली को उपलब्ध कराए जाने के सामान्य लक्ष्य को नि
र्धारित किया जाना चाहिए और शिक्षा प्राप्ति के रास्ते को सुगम व सुलभ्य बनाए जाने की कोशिश की जानी चाहिए।
>>>
========== RESTART: E:\Desktop\nlp1.py ==========
```

Fig. 4 Paragraph after 40 percent compression ratio.

RESULT AND ANALYSIS

To test our summarization, we collected various Hindi documents. The documents are typed and saved in the text files using UTF-8 format. It is very difficult to determine whether a summary is good or bad. The summary evaluation methods can be broadly categorized as human evaluation methods and automatic (machine-based) evaluation methods. A human evaluation is done by comparing system-generated summaries with reference/model summaries by human judges. The automatic evaluations may lack the linguistic skills and emotional perspective that a human has. Hence although automatic evaluation is not perfect compared to the human evaluation, it is popular primarily because the evaluation process is quick even if summaries to be evaluated are large in number. Since automatic evaluation is performed by a machine, it follows a fixed logic and always produces the same result on a given summary.

CONCLUSION AND FUTUTE SCOPE

Text Summarization is increasing as a sub-branch of NLP as a demand for compressive, substantive, abstract subject due to a large amount of knowledge on the net. This is a single document text summarization method for Hindi. Many techniques have been developed for summarizing English text(s). But, a very few attempts have been made for Hindi text summarization. The most important advantage of using a text summarization is that it reduces the reading time.

The performance of the proposed system may further be improved by improving stemming process, exploring more number of statical and linguistic features like proper noun and applying learning algorithm for effective feature combination. In future, more features like named entity recognition, cue words, context information, world knowledge etc, can be added to improvise the technique. Also, same technique can be applied on various domains. It can also be extended to work on multiple documents.

REFERENCES

[1] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159–165.

[2] Student, P., & COE, D. M. (2015). A comparative study of hindi text summarization techniques: Genetic algorithm and neural network.

[3] Wikipedia contributors. (2018). Automatic summarization — Wikipedia, the free encyclopedia. Retrieved from https://en.wikipedia.org/w/index.php?title= Automatic summarization&oldid=822496672 ([Online; accessed 29-April2018])

[4] 1Ajinkya Zadbuke, 2 Sahil Pimenta, 3Deepen Padwal, 4Varsha Wangikar, "Automatic Summarization of News Articles using TextRank", Volume 6, Issue 3, March 2016

[5] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM (JACM), 16(2), 264–285.