Automatic Textual Content Summarization using Natural Language Processing

Anjana KUmari[1], Raju Manjhi[2], Arnav Raviraj[3] Dr. Ajay Kumar Murari[4], Dr. I.Mukherjee[5]

[1]Research Scholar, [2]Marwari College Ranchi c[3]Systems Engineer, Infosys
[4]Associate professor,Department of Physics,Vinovabhave University, hazaribagh
[5]Professor BIT Mishra Ranchi

ABSTRACT

This paper explores on sentence extraction based single Document rundown. It saves time in our every day work once we get summed up information. Today there are such countless Documents, articles, papers and reports accessible in advanced structure, however the greater part of them need outlines. Programmed text Synopsis is a procedure where a PC sums up a text. A text is given to the PC and the PC returns a necessary concentrate of the first text archive. Our strategies on the sentence extraction-based message synopsis task utilize the chart based calculation to compute significance of each sentence in archive and most significant sentences are removed to create report outline. These extraction based message outline strategies give an ordering weight to the record terms to process the closeness esteems between sentences

Keywords

NLP, Summarization, Sentence Ranking

1. INTRODUCTION

Today internet contains vast amount of electronic collections that often contain high quality information. However, usually the Internet provides more information than what is needed. User wants to select best collection of data for particular information need in minimum possible time .Text summarization is one of the applications of information retrieval, which is the method of condensing the input text into a shorter version, preserving its information content and overall meaning. World Wide Web contains vast amount of information which contain a sort of information which may not be useful. Many technologies are used to get data required which has given rise to technique called Summarization. A summary can be defined as a text that is produced from one or more texts, which contains a significant portion of the original text, which is summarization given data. Text Summarization is a method of getting an output as of extract document. When this is done by means of computer, i.e. automatically, we call this as Automatic Text Summarization. Our paper addresses you the techniques used for summarization.

2. FEATURE BASED TEXT SUMMARIZATION

This paper describes that the last six decades, the problem of text summarization has been approached from many different perspectives, in various domains and using various paradigms. Some of the techniques used to for text summarizer include

3. CONNECTIONIST APPROACH TO GENERIC TEXT SUMMARIZATION

The aim here is to auto summarizes large documents. This approach utilizes adaptive, incremental learning and knowledge representation system that evolves its structure and functionality. This approach proposes usage of Part of Speech disambiguation using a recurrent neural network, a paradigm capable of dealing with sequential data.

Basically summary can be of two types Extractive and Abstractive. Abstractive summary represents use of Natural Language Processing (NLP) whereas Extractive summary is based on copying exact sentences from source document.

Ranking Of Text Units According To Shallow Linguistic Features: This approach recognizes the most prominent text/sentences using various shallow linguistic features, taking degree of connectedness among the text units into consideration so that it minimizes the poor linking sentences in the resulting text summary. This method highlights theeffect of lexical chain scoring after the nouns and compound nouns are chained by searching for lexically organized relationships between words in the text using WorldNet and using lexicographical relationships such as synonyms and hyponyms. All the sentences are ranked or given preferences on the basis of the sum of the scores of the words in each sentence in order to extract a summary. The scores of words are decided using various features like term frequencies, cue words and phrases, measuring lexical resemblance (measuringchain score, word score and finally sentence score)etc.

4. PROPOSED SYSTEM

Today internet contains vast amount of electronic collections that often contain high quality information. However, usually the Internet provides more information than is needed. User wants to select best collection of data for particular information need in minimum possible time. Text summarization is one of the applications of information retrieval, which is the method of condensing the input text into a shorter version, preserving its information content and overall meaning. There has been a huge amount of work on query specific summarization of documents using similarity measure. This paper focuses on sentence extraction based single document summarization. Our propose method works on the sentence extraction-based text summarization task use the graph based algorithm to calculate importance of each sentence in document and most important sentences are extracted to generate document summary. These extraction based text summarization methods give an indexing weight to the document terms to compute the similarity values between sentences.

Summarization is a NP hard problem of Processing because, to do it properly, one has to really understand the point of a text. This requires semantic analysis, discourse processing, and inferential interpretation (grouping of the content using world knowledge). To achieve the best to the art ofsummarization we have decided the following objective for our project.

i. **Preprocessing**
Parse the document and generate sentences.

ii. **Graph Building**
This represents a sentence as a node with all its properties and methods to handle with its behavior.

iii. **Sentence Ranking Algorithm**
The basic approach of Sentence Rank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally. First of all

A document ranks high in terms of Sentence Rank, if other high ranking documents link to it.

The original Sentence Rank algorithm was described by Lawrence Page and Sergey Brin in several publications. It is given by

$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$

Where

PR(A) is the Sentence Rank of page A,

PR(Ti) is the Sentence Rank of pages Ti which link to page A,

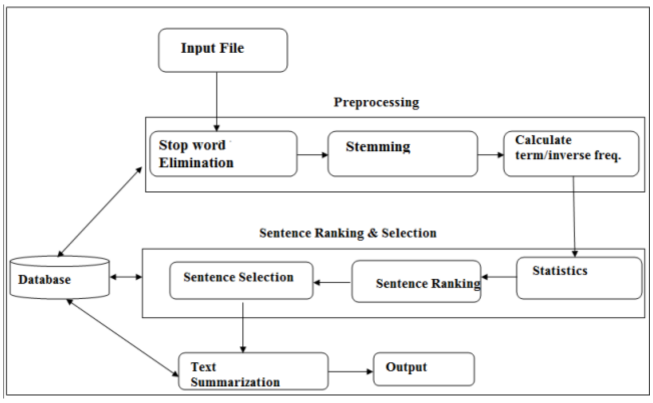
C(Ti) is the number of outbound links on page Ti and

d is a damping factor which can be set between 0 and 1.

iv. **Summarization**
The output of the Project will be Text summarized data.

5. DESIGN AND IMPLEMENTATION CONSTRAINTS

1. If there is power failure system cannot recover the ongoing session.
2. Time: The project must be completed in a time span of 5 months including testing and Documentation.
3. The system must provide accurate results i.e. it must execute the commands properly and must
4. Perform the desired functions.
5. 5. The system must be fast in processing the request and sending the appropriate response.
6. Provides better flexibility.



System Architecture

Advantages

1. Relatively fast (compared to full parse)
2. 2 Provides a good general idea or feel for content.
3. Can do multiple-document summaries.

Disadvantages

1. Does poorly when document doesn't contain good summary sentences.

6. CONCLUSION

The vast growth in the rate of information due to internet has called for a need of efficient summarization systems. Although the research on text summarization has started so many years ago, there is still a long trail to walk and some more things to be researched as well. This literature explores the recent trend in summarization system that comes from novice procedure to this time of computer, where natural language processing is used to generate the summary resemble with human expert. It is concluded that the achieved results of Summarization Ranking Module are a promising start toward further studies. Future researches in this field would mainly concentrate on the ability to find efficient ways of automatically evaluating these systems and on the development of measures that are objective enough to be commonly accepted by the research community.

7. REFERENCES

[1] R. S. Prasad, U. V. Kulkarni, J. R. Prasad, “A Novel Evolutionary Connectionist Text Summarizer (ECTS),”, 2009, *IEEE Xplore*.

[2] Pankaj Gupta, Vijay Shankar Pendluri, Ishant Vats, “Summarizing text by ranking text units according to shallow linguistic features”, Feb. 13~16, 2011 ICACT, 2011.

[3] Rajesh Shardanand Prasad, Uday. V. Kulkarni, “Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization”, Journal of Computer Science 6 (11): 1366- 1376, 2010 ISSN 1549-3636, 2010 Science Publications.

[4] Uplavikar Nitish Milind, Wakhare Sanket Shantilalsa, Prof. Dr. R.S. Prasad, "International Journal of Advances in Computing and Information Researche ISSN: 2277-4068, Volume 1– No.2, April 2012"