iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** V **Month of publication:** May 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Implementation of Automatic Text Summarization

Aarti Patil[1], Komal Pharande[2], Dipali Nale[3], Roshani Agrawal[4], S.P.Bunjkar[5]
Sinhgad Institute of Technology & Sciences, Narhe, Pune, India

Abstract*— This paper investigates on sentence extraction based single Document summarization. It saves our time in daily work once we get summarized data. Today there are so many reports, Documents, papers, and articles available in digital form, but most of them lack summaries. Automatic text Summarization is a technique where a computer summarizes a text. A text is given to the computer and the computer returns a required extract of the original text document. Our methods 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. Then the clustering of documents is done as per the domain of the documents, along with it labels are given to the clusters.

Keywords —NLP, summarization, Graph building page ranking, Heuristics and Statistics, clustering, clustering labels

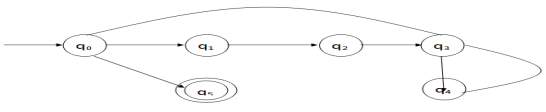
I. INTRODUCTION

Today internet contains vast amount of electronic collection 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 from, preserving its information content and overall meaning. Now-a-days internet is only major source of information used around the world. World Wide Web contains vast amount of information which contain a sort of information which may not be useful for particular purpose. Many technologies are used to get desired data which has given rise to technique called Summarization. Our paper addresses you the techniques used for summarization. Internet contain information is form of text, audio, video, images and any other format. It's difficult to summarize document which consist of this entire format. Many techniques have been used to get rid of this problem. Text Summarization is a method of getting a document which contains significant portion of original text. 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 the effect 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 (measuring chain score, word score and finally sentence score)etc.

II. MATHEMATICAL MODEL

Let S be a closed Document Summary system such that S = {D,SP,SD }
Where
D represents the Input Document,
SP is Summary Percentage
SD is Summarized Document
Activity1 Let fe be a rule of S,I into A such that for given Summary; it returns fe(D,SP) |SD.

Finite State Machine



FSM

Fig::1FSM

P-NP COMPLETE

- Problem Type

K ∈ NP Set
The required Time : O(n*n)

- This problem are : P

K ∈ {x,y} //x=input function,y=output function
K ∈ P
Is require time : O(n * n)
So that we can solve this problem by
Type = P
P= NP

III. PREVIOUS WORK

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.

IV. 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. Here user can specify the number of lines so that user can get the summary of that data into desired number of lines, as we are ranking the statements according to the priority. As user enter n number of lines to get output then first n lines are printed as output. Our is the NP Hard problem. To obtain best solution we apply the following steps:

A. Preprocessing

Parse the document and generate sentences.

B. Graph Building

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

C. 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.

D. Summarization

The output of the Project will be Text summarized data.

E. clustering

Now the summarized data will be clustered under their domain. The name of the domain will be decided by system itself. For that we are using Lingo Algorithm. Lingo algorithm better than k-means at reconstruction and extension-over-time. The key characteristic of the Lingo algorithm is that it reverses the traditional clustering pipeline: it first identifies cluster labels and only then assigns documents to the labels to form final clusters. Cluster labels are longer and more descriptive in Lingo

Algorithm.

1) Design and Implementation Constraints

a) If there is power failure system cannot recover the ongoing session.
b) Time: The project must be completed in a time span of 5 months including testing and Documentation.
c) The system must be fast in processing the request and sending the appropriate response.
d) Provides better flexibility.

2) Assumption and Dependencies

a) The System is centralized
b) We assume that the NLP API will be used for developing application.
c) Single Document at a time can be summarized.
d) Document should be in the form of TXT or RTF format only.
e) Test challenges the assumptions, risks and uncertainty inherent the work of the other disciplines, addressing those concerns by concrete demonstration and impartial evaluation.

3) Advantages

a) Can specify number of lines of output
b) Relatively fast (compared to full parse)
c) Provides a good general idea or feel for content.
d) Can do multiple-document summaries.

V. FUTURE SCOPE

A. Multiple document summarize can be achieved
B. It can be further expanded for multiple languages.

VI. CONCLUSION

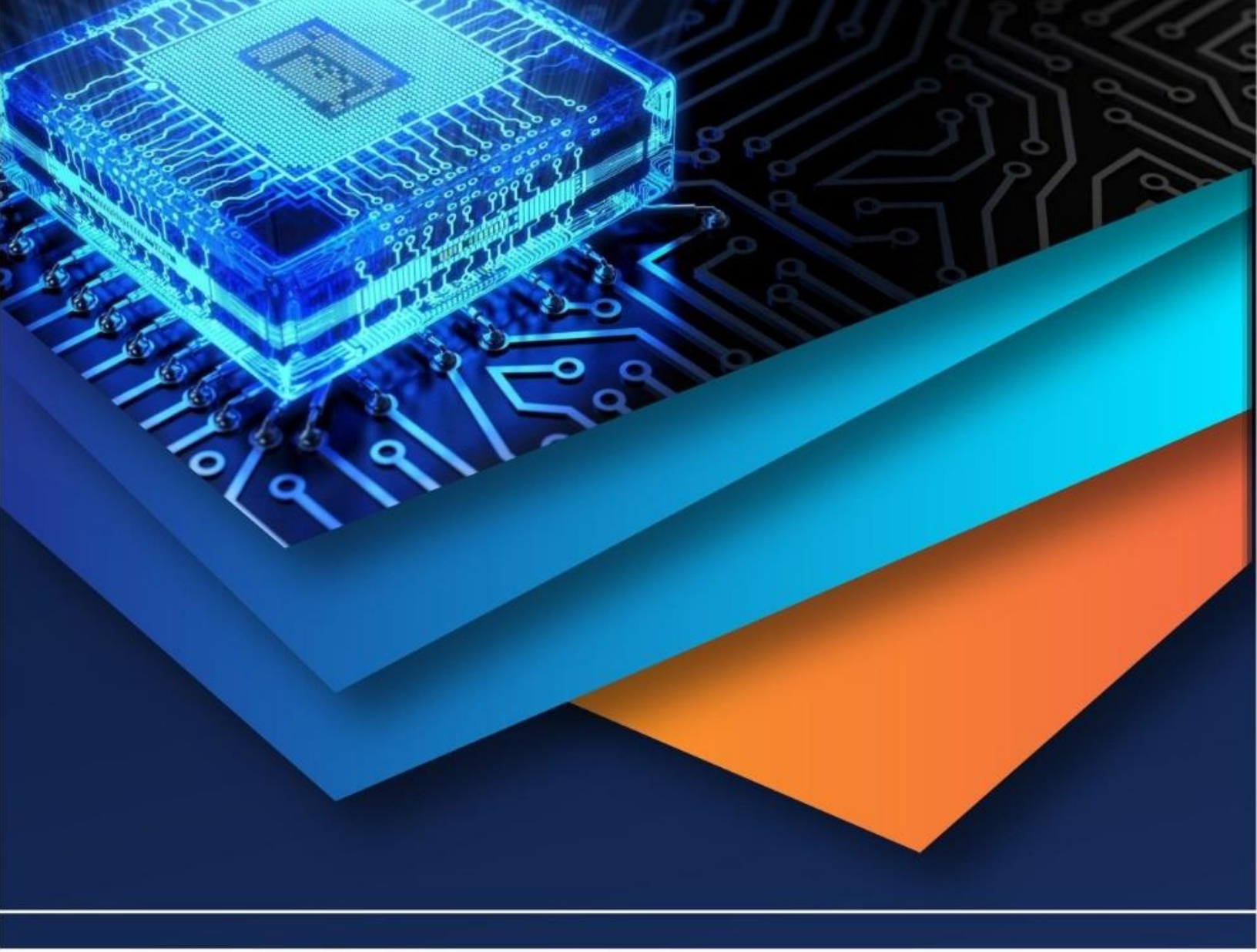
The vast growth in the rate of information due to internet has called for a need of efficient summarization systems. Although the research on text summarization has started so many years ago, there is still a long trail to walk and some more things to be researched as well. This literature explores the 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.

VII. ACKNOWLEDGMENT

We would like to thanks our guide Prof.S.P.Bunjkar madam for her unconditional support.

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"



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY