International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 03 (March 2015), PP.57-59

A Survey of Various Methods for Text Summarization

Vipul Dalal[1], Yogita Shelar[2]
[1]*Computer Department, Vidylakar Institue Of Technology, Wadala(w).*
[2]*Student of Information Technology, Vidylakar Institue Of Technology, Wadala(w).*

Abstract:- Document summarization means retrieved short and important text from the source document. In this paper, we studied various techniques. Plenty of techniques have been developed on English summarization and other Indian languages but very less efforts have been taken for Hindi language. Here, we discusses various techniques in which so many features are included such as time and memory consumption, efficiency, accuracy, ambiguity, redundancy.

Keywords:- summarization, redundancy, efficiency, ambiguity, accuracy.

I. INTRODUCTION

Document summarization refers to the task of creating document surrogates that are smaller in size but retain various characteristics of the original document. To automate the process of abstracting, researchers generally rely on a two phase process. First, key textual elements, e.g., keywords, clauses, sentences, or paragraphs are extracted from text using linguistic and statistical analyses. In the second step, the extracted text may be used as a summary. Such summaries are referred to as 'extracts'. Alternatively, textual elements can be used to generate new text, similar to the human authored abstract. Summarization of Hindi documents contains historical information is also plays as important role for students and teachers who want to read a large number of documents related to history. Summarization system helps them to read and learn the shorter version of overall complete document Summarization system helps them to read and learn the shorter version of overall complete document..

Automatic Text Summarization is an important and challenging area of Natural Language Processing. The task of a text summarizer is to produce a synopsis of any document or a set of documents submitted to it Analysis of Text-Documents has been an active area of research for the past few years. It involves extensive use of Natural Language Processing techniques for analysing semantic structures of the text. Semantic analysis of a document means to analyse the meaning or transitions in meaning of the sentences or of different clauses and the relation among them. There are a number of approaches to semantic analysis. Semantic analysis can be done at the sentence level, the paragraph level, or even at the text level on different languages.

Hindi is the official and the most widely spoken language in India. As of pronoun resolution, for the pronouns having more than one possible antecedent, the pronoun resolution mechanism of this system captures the ambiguity. The approach discussed here is to perform semantic analysis at the sentence level where the Hindi text is scanned for pronouns and the corresponding referents resolved.

Summaries differ in several ways. A summary can be an extract i.e. certain portions (sentences or phrases) of the text is lifted and reproduced verbatim, whereas producing an abstract involves breaking down of the text into a number of different key ideas, fusion of specific ideas to get more general ones, and then generation of new sentences dealing with these new general ideas.

II. RELATED WORK

Various methods have been proposed to achieve extractive summarization. Most of them are based on scoring of the sentences.

Dr.Latesh Malik, et. al.[1], Discussed single document summarization using extraction method for Hindi text, which uses statistical and linguistic features. It uses Hindi Wordnet to tag appropriate POS of word for checking SOV of the sentences which uses sixstatistical and two linguistic features. It also uses genetic algorithm to optimize the summary generated based on the text feature terms with less redundancy.

Ibrahim F. Moawad, et. al.[2], Described a noval approach is presented to create an abstractive summary for a single document using a rich semantic graph reducing technique. The approach summaries the input document by creating a semantic graph called Rich Semantic Graph for the original document, reducing

the generated semantic graph to more abstracted graph, and generating the abstractive summary from the reduced graph but in English.

Sachin Agarwal, et. al.[3], Proposed the algorithm for anaphora resolution has been tested extensively. The accuracy of anaphora resolution is 96% for simple sentence not for original document and complex sentences; the accuracy is of the order of 80%. This method works by searching sentences in the text that are semantically related thorough anaphors, analyzing their semantic s and exploiting the latter for s resolving respective anaphors.

Ng Choon-Ching, et. al.[4], Proposed an existing need for text summarizers that small devices like PDA has emerged the development of text summarization of web pages. Authors have identified problems for text summarization in several areas such as dynamic content of web pages, diverse summary definition of text, and different benchmark of evaluation measurements. Besides, authors also found advantages of certain methods that increased the accuracy of web page classification. In the future work, author plan to investigate machine learning techniques to incorporate additional features for the improvement of text summarization quality. The additional features authors are currently considering include linguistic features such as discourse structure, lexical chains, semantic features such as name entities, time, location information etc

Visual Gupta, et. al.[5], Describe the Punjabi text extractive system which consist of two phases 1) Pre Processing 2) Processing. In this paper term pre processing is defined as the phase which identify the word boundary, sentence boundary, Punjabi stop words elimination etc. and the processing phase sentence features are calculated and a weight is assigned to each sentence on the reference of which unwanted sentences are eliminated from the input text. It is described that the author tested the proposed system over fifty Punjabi news documents (with 6185 sentences and 72689 words) from Punjabi Ajit news paper and fifty Punjabi stories (with 17538 sentences and178400 words). Accuracy of the system is varies from 81% to 92 %.

Niladri Chatterjee, et. al.[6], Described summarization technique for text document exploiting the semantic similarity between sentences to remove the redundancy from the text. It uses Random Indexing for compute the semantic similarity scores of sentences and graph-based ranking algorithms have been employed to produce an extract of the given text. The important is that the problem of high dimensionality of the semantic space corresponding text should be tackled by random indexing which is less expensive in computations and memory consumption and it included a training algorithm using Random Indexing which has to construct the Word space on complied text document so that resolve the ambiguities such as more efficiency.

M. C. Padma, et. al.[7], In a multi-script multi-lingual environment, a document may contain text lines in more than one script/language forms. It is necessary to identify different script regions of the document in order to feed the document to the OCRs of individual language. With this context, this paper proposes to develop a homothetic algorithmic model to identify and separate text lines Telugu, Hindi and English scripts from a printed multilingual document. The proposed method uses the distinct features of the target script and searches for the text lines that possess the anticipated features. Experimentation conducted involved 1500 text lines for learning and 900 text lines for testing. The performance has turned out to be 98.5%.

Erika Velazquez-Garcia, et. al.[8], Proposed A novel method to organize, search and display groups of document according to topics they contain based on the collection of synonyms, and hypernyms, hyponyms of each terms Thus, each user would have a personalized and dynamic organized of documents thereby it takes more time for text processing.

Sunil Kumar, et. al.[9], Suggested a novel scheme for the extraction of textual areas of an image using globally matched wavelet filters. A clustering-based technique has been devised for estimating globally matched wavelet filters using a collection of ground truth images. We have extended our text extraction scheme for the segmentation of document images into text, background, and picture components (which include graphics and continuous tone images). Multiple, two-class Fisher classifiers have been used for this purpose. We also exploit contextual information by using a Markov random field formulation-based pixel labeling scheme for refinement of the segmentation results. Experimental results have established effectiveness of our approach..

M. Swamy Das, et. al.[10], Described document should be composed of text contents in different languages in multilingual country. It is necessary to identify the language region of the document before feeding the document to the related OCR system. Advantage of this paper is that a model to identify script type of different text portions using visual clues. Here seven features are covered, such as, bottom max row, top

horizontal lines, vertical lines, bottom component, tick component and top holes, and bottom holes have been used to identify the script document. Identification accuracy of above 93%is achieved.

III. CONCLUSIONS

Hindi is the official and the most widely spoken language in India. In this paper, we discussed various methods for summarization. But many of techniques are found on English and other languages but very few methods on Hindi text document. Summarization of Hindi documents contains historical information is also plays as important role for students and teachers who want to read a large number of documents related to history. Summarization can be two types: 1. Extractive Summarization 2. Abstractive Summarization. In both extractive and abstractive summarization technique rule based approach can be used in which various handcrafted rules are to be created on the basis of which summary of the text document can be generated.

ACKNOWLEDGMENT

I would like to express sincere thanks to Mr. Vipul Dalal, who has given us the new vision to think on "Text Summarization" with different angles pertaining to problems area actually being faced by the technical experts while execution of works.

I take this opportunity to thank Mr. Vipul Dalal for his encouraging words & valuable time enabling me to come out with useful knowledge material.

REFERENCES

[1]. Dr.Latesh Malik,"Test Model for Summarizing Hindi Text using Extraction Method",(Proceedings of 2013 IEEE Conference on Information and Communication Technologies) (ICT 2013).

[2]. Ibrahim F. Moawad, Information Systems Dept.Faculty of Computer and Information Sciences "Semantic Graph Reduction Approach for Abstractive Text Summarization",(Ain shams University Cairo, Egypt ibrahim_moawad@cis.asu.edu.eg 2012 IEEE).

[3]. Sachin AGARWAL Manaj SRIVASTAVA, "Anaphora Resolutio88888888n in Hindi Documents",(Indian Institute of Information Technology – Allahabad Allahabad, UP, India2007IEEE)

[4]. Do Phuc, University of Information Technology,"Using SOM based Graph Clustering for Extracting Main Ideas from Documents",(VNU-HCM HoChiMinh City,VietNam phucdo@uit.edu.vn 2008 IEEE) Vishal Gupta and Gurpreet Singh Lehal," Automatic Punjabi Text Extractive Summarization system."

[5]. Proceedings of COLING 2012: Demonstration Papers, pages 199–206, COLING 2012, Mumbai, December 2012.

[6]. Niladri Chatterjee,"Extraction-Based Single-Document Summarization Using Random Indexing",(19th IEEE International Conference on Tools with Artificial Intelligence IEEE2007).

[7]. M. C. Padma, P. A. Vijaya, "Monothetic Separation of Telugu, Hindi and English Text Lines from a Multi Script Document", (Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009)

[8]. Erika Velazquez-Garcia, Ivan Lopez-Arevalo, Victor Jesus Sosa-Sosa Information Technology, Laboratory CINVESTAV – Tamaulipa,"Representing Document Semantics by Means of Graphs", (http://www.google.com visited in September 2011).

[9]. Sunil Kumar, Rajat Gupta, Nitin Khanna, Student Member, IEEE, Santanu Chaudhury, and Shiv Dutt Joshi, "Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model", (IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 16, NO. 8, AUGUST 2007).

[10]. M. Swamy Das, D. Sandhya Rani, C R K Reddy, "Heuristic based Script Identification from Multilingual Text Documents",International Conf. On Recect Advances in Information Technology (RAIT-2012).