Gist Generation of Hindi Document Using Statistical Method
A REVIEW PAPER

Mrs. A.N.Pimpalshende
Computer Technology department
Priyadarshini Institute of engineering and Technology,Nagpur
Nagpur,India
anjuraut@rediffmail.com

Dr. A.R. Mahajan
Computer Science and Engineering department
Priyadarshini Institute of engineering and Technology,Nagpur
Nagpur,India
armahajan@rediffmail.com

***Abstract*—**For the blessing of World Wide Web, the corpus of online information is gigantic in its volume. Search engines retrieve specific information from this huge amount of data. But the outcome of search engine is unable to provide expected result as the quantity of information is increasing enormously day by day and the findings are abundant. So, the automatic text summarization is demanded for salient information retrieval in short time. Text summarization is to compress an original document into a shortened version by extracting the most important information out of the document. Many approaches use statistics and machine learning techniques to extract sentences from documents. An extractive summarization method consists of selecting important Sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. In the absence of natural language understanding system, it is required to design an appropriate system. Gist generation is a difficult task because it requires both maximizing text content in short summary and maintains grammaticality of the text. In this paper we propose a statistical approach to generate a gist of a Hindi document.

Keywords- *Text Summarization, extractive summary, Natural language understanding* ***(key words)***

I. INTRODUCTION

As the information resources in both online and offline are increasing exponentially, the major challenge is to find relevant information from large amount of data in short time. Text summarization is to compress an original document into a shortened version by extracting the most important information out of the document. Instead of keyword extraction, text summarization identifies the most important paragraphs or sentences from a given document, excluding unimportant detailed information [1]. The summaries serve as quick guide to interesting information, providing a short form for each document in the document set: reading summary makes decision about reading the whole document or not, it also serves as time saver.

Summaries in Everyday life:

Headline	summaries of newspaper articles
Biography	resume, obituary
Digest	summary of stories on the same topic
Highlight	Summary of an event (meeting, sport event, etc.)
Abstract	summary of a technical paper
Bulletin	weather forecast, stock market, news
Trailer	Movie, speech

The aim of this work is to propose the system for generating the text summary of Hindi language document using statistical method.

II. APPROACHES OF TEXT SUMMARIZATION

A. Text Summaries are of two types: Extractive summaries and Abstractive summaries

a) **Extractive summary**: Extractive summaries involve extracting relevant sentences from the source text in proper order. The relevant sentences are extracted by applying statistical and language dependent features to the input text.

In most of cases in the world we prefer to make extractive text summaries due to its ease in generating text summary. Difficulties of extractive text summary [1] [2] are: 1) As compared to average summaries, extractive summaries are normally lengthy because certain sections of text which are not required in summary may also be included in it. 2) In many cases essential information is usually present across different lines, and usually extractive summaries may not collect it unless it is lengthy enough for covering all these lines

b) **Abstractive Summary**: Abstractive text summaries are made by applying natural language understanding. Human beings usually make summaries in abstractive way. Moreover abstractive summaries can also involve the words or sentences which are not present in the input text. Abstractive text summarization is difficult because as compared to human beings, computers have limited capabilities of language understanding, so alternative methods must be considered

B. Extractive based summarization techniques:

a) Statistical Method: Text summarization based on this approach relies on the statistical distribution of certain features and it is done without understanding whole document. It uses classification and information retrieval techniques. Classification methods classify the sentences that can be part of the summary depending on the training of data. Information

retrieval technique uses Position, length of sentence or word occurrences in the document. This method extracts sentences that occurs in the source text, without taking into consideration the semantics of the words

b) Linguistic Method:

In this, method needs to be aware of and know deeply the linguistic knowledge, so that the computer will be able to analyze the sentences and then decide which sentence to be selected. It identifies term relationship in the document through part-of-speech tagging, grammar analysis, thesaurus usage and extract meaningful sentences. Parameters can be cue words , Title feature or Noun and verbs in the sentences[3] In practice, linguistic approaches also adopt simple statistical computation (term-frequency-inverse-document-frequency (TF-IDF) weighting scheme) to filter terms.

III. RELATED RESEARCH OVERVIEW

A lot of research in single document summarization has gone into finding out the relevant segments from the text, ranking them and finally generating the summary which expresses most of the important points. The task of gist generation is strongly connected to traditional text summarization [3] and emphasizes the extractive approach which selects words, sentences or paragraphs from the document to provide a summary. In this section, we are presenting the different methods used for extractive text summarization in Indian languages.

Vishal Gupta [4] presented "Punjabi Text Summarization" in this Punjabi Keywords selection feature (TF-ISF approach) and number feature are used for sentence selection. Mathematical regression is used to estimate the text feature weights based on fuzzy scores of sentences of 50 Punjabi news documents for summary generation.

Kumar and Devi[5] presented "Tamil language summarization system " in this Scoring of sentences are based on graph theoretic scoring technique for summary generation.

CDAC (Centre for development of advance computing) Noida. Presents the system "Automatic text summarization software for Hindi text" in this system summary is generated using Statistics based technique, language oriented & heuristic technique had been applied.

Islam and Masum presents "corpus oriented text summarization system 'Bhasa' for Bengali language" in this scoring the files of corpus in which query words are having highest frequency and then producing the summary of text documents on the basis of query words by applying vector-space-term-weighting.

Sarkar (2012) proposed Bengali text summarization by sentence extraction and has investigated the impact of thematic term feature and position feature on Bengali text summarization. The proposed summarization method is extraction based. It has three major steps: (1)preprocessing (2) sentence ranking (3) summary generation. The preprocessing step includes stop-word removal, stemming and breaking the input document in to a collection of sentences. After an input document is of a sentence is computed in such a way that the first sentence of a document gets the highest score and the last sentence gets the lowest score. Long sentences are given preference in summary A summary is produced after ranking the sentences based on their scores and selecting K-top ranked sentences, when the value of K is set by the user. To increase the readability of the summary, the sentences in the summary are reordered based on their appearances in the original text.[5]

Sarkar (2012) proposed another approach for summarizing Bengali news documents. It describes a system that produces extractive summaries of Bengali news documents. The ultimate objective of produced summaries is defined as helping readers to determine whether they would be interested in reading a particular document. To this end, the summary aims to provide a reader with an idea about the theme of a document without revealing the in-depth detail. The approach presented here has four major steps (1) preprocessing (2) extraction of candidate summary sentences (3) ranking the candidate summary sentences (4) summary generation.

Kallimani et al. (2012) proposed a new technique for summarizing the longer text documents by considering one of the South Indian regional languages (Kannada). It deals with a single document summarization based on statistical approach. The purpose of summary of an article is to facilitate the quick and accurate identification of the topic of the published document. The objective is to save prospective readers' time and effort in finding the useful information in a given huge article.

IV. LANGUAGE CHARACTERSTICS

Hindi is the national language of India. It is one of several languages spoken in different parts of the sub-continent. National should be understood as meaning the official or link language. The homeland of Hindi is in the North of India, but it is studied, taught, spoken and understood widely throughout the sub-continent, whether as mother tongue or as a second or a third language. Hindi is written in Devanagari script. The script is phonetic, so that Hindi, unlike English, is pronounced as it is written. Therefore, it is to learn the characters of the script and the sounds of the language at the same time. There are 33 consonants and 11 vowels in Hindi. Additionally, there are also many conjunct consonants. Hindi consonants are divided into groups on the basis of phonetic properties of their formations such as plosives, nasals, fricatives, flapped and tapped sounds, and semi vowels. Each Devanagari script character represents a syllable, not the alphabet. English preposition-like Hindi words or "postpositions" follow their related words. Hindi sentence syntax follows subject-object-verb structure.

V. PROPOSED SYSTEM.

The goal of the system is to extract most relevant sentences from the Hindi document. The proposed method uses statistical method for sentence generation. Summarization System consists of 3 major steps. Preprocessing, extraction of feature terms and ranking of sentences based on optimized feature weight.

4.1. Preprocessing: is a structured representation of the original documents.

4.11 Sentence segmentation: Decompose sentences along with its word count.

4.12 Tokenization is a process of splitting of sentences into words by identifying spaces, comma.

4.13 Stop word removal: remove common words with no semantics in English a, an,and in hindi "par" "en", "unhe"….
4.14 Stemming is used to check similarity feature. In this obtain the stem or root of the word or obtain similar words

4.2 Processing: Every sentence is represented by a vector of feature terms. Which checks for every sentence statistically and linguistically? Each sentence has a score based on the weight of feature terms which is used for sentence ranking. Top ranked sentences are selected for final summary.
Different Features for sentence selections are.

1. Average term frequency.
2. Sentence length
3. Sentence position
4. Numeric data
5. Sentence to Sentence similarity
6. Title score
7. SOV qualification
8. Subject similarity
9. Proper noun
10. Upper-case word
11. Cue phrase feature

4.3 Sentence generation: assigns a score to each sentence and then rank the sentences according to their scores. The sentence with higher scores are included in the summary depending on compression ratio. In proposed method we are using swarm intelligence approach for selection of important sentences to include in the summary. These sentences are arranged in grammatical order.

Automatic generated summaries can be evaluated using following parameters.
Precision: It evaluates correctness for the sentences in the summary.

$$P = \frac{\text{Retrieved sentences } \Omega \text{ Relevant sentence}}{\text{Retrieved sentences}}$$

Recall: It evaluates proportion of relevance included in the summary.

$$R = \frac{\text{Retrieved sentences } \Omega \text{ Relevant sentences}}{\text{Relevant sentences}}$$

F1 score: The F1 score is a standard way to mix the two numbers in a single score

V I. CONCLUSION

Many automatic text summarization systems are commercially or non-commercially available for most of the commonly used natural languages for English and other foreign languages, but when it comes to Indian languages, automatic text summarization systems are still lacking. But now a day's lot of research is going on for Indian regional languages. This is a proposed system to generate text summary for Hindi text document

REFERENCES

[1] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques," InternationalJournal of Emerging Technologies in Web Intelligence, vol.2. , pp. 258-268, 2010)

[2] K. Sarkar, "Bengali text summarization by sentence extraction," In Proceedings of International Conference on Business and Information Management(ICBIM-2012), NIT Durgapur, pp. 233-245, 2012.

[3] K. Sarkar, "An approach to summarizing Bengali news documents," In proceedings of the International Conference on Advances in Computing, Communications and Informatics, ACM, pp. 857-862, 2012.

[4] T. Islam and S. M. A. Masum, "Bhasa: A Corpus Based Information Retrieval and Summarizer for Bengali Text," Macquarie University, Sydney, Australia, 2004.

[5] V. Gupta and G.S. Lehal, "Automatic Text Summarization for Punjabi Language," *International Journal of Emerging Technologies in Web Intelligence*, vol. 5, pp. 257-271, 2013.

[6] V. Gupta and G. S. Lehal, "Automatic Punjabi Text Extractive Summarization System," *International Conference on Computational Linguistics COLING '12*, IIT Bombay, India, pp. 191-198, 2012.

[7] U. Hahn and I. Mani, "*The challenges of automatic summarization*," IEEE Computer, vol. 33(11) pp. 29–36, November 2000

[8] Chen, Kesong Han and Guilin Chen," *An Approach to sentence selection based text summarization*", In Proceedings of IEEE TENCON02, pp489-493, 2002