INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A REVIEW PAPER ON TEXT SUMMARIZATION OF HINDI DOCUMENTS

Nitika Jhatta[1], Ashok Kumar Bathla[2]

[1]M.tech. Student,CE Dept.,Yadavindra College of Engineering, Talwandi Sabo,Punjab,India.nitikajhatta12@gmail.com
[2] Assistant Professor,CE Dept.,Yadavindra College of Engineering,Talwandi Sabo,Punjab,India. ashokashok81@gmail.com

Abstract: - A summary of a document is a (much) shorter text conveys the most important information from the source document. Summary of the text must contain important information from the document. Summary of the text can be generated from a single document or from multiple documents. In single-document summarization, the summary of only one document is to be built, while in multi-document summarization the summary of a whole collection of documents (such as all today's news or all search results for a query) is built. In this paper authors have represented review on single-document summaries for the text data written in the various Languages. This paper also presents detection and removal of Deadwood, Extractive and Abstractive summarization methods for Punjabi language. An Extractive summarization method only decides, for each sentence, whether or not it will be included in the summary. An Abstractive summarization process consists of "understanding" the original text and "re-telling" it in fewer words.

Keywords*:* Deadwood, Extractive, Abstractive, Summarization

1. Introduction

an automatically built summary in its list of retrieval results, for the user to quickly decide which documents are interesting and worth opening for a closer look—this is what Google models do some degree with the snippets shown in its search results. Other examples include automatic construction of summaries of news articles or email messages to be sent to mobile devices as SMS; summarization of information for government officials, businessmen, researches, etc., and summarization of web pages to be shown on the screen of a mobile device, among many other. Summarization of Hindi documents contain 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. Historical documents in Hindi language contain information which consists of important dates with their associated events 1947, the year in which there are a number of scenarios where automatic construction of such summaries is useful. For example, an information retrieval system could present India got freedom: , names of famous persons like "Bhagat Singh" , information that contain names of places like "various battles held in

Panipat" etc. Hindi language is also used in Government Officials. The Constitution of India designates a bilingual approach for official language of the Government of India employing usage of Hindi written in devanagri scripts well as English. Hindi find everyday use for important official purposes such as parliamentary proceedings, judiciary, communication between central government and state government. The number of native Hindi speakers range between 14.5 to 24.5 %. In total India population however, other dialects of Hindi termed as Hindi languages are spoken by 45% of Indians mostly accumulated from states falling under the Hindi belt.

Summarization can be two types:

1. Extractive Summarization

2. Abstractive Summarization

Extractive Summarization:

An extractive summarization method only decides, for each sentence, whether or not it will be included In the summary. In extractive summarization system different weights are assigned to each sentence of the document on which sentence is selected to get added for the summary. Weights can be assigned to the sentences according to the position of the sentence in the document i.e. sentences in the beginning and at the end are assigned more weight as they are supposed to contain more valuable information. Weights can also be assigned according to the type of information they contain. For example if a sentence contain name of person, date of the event occurred then more weight is given to that sentence than those which do not contain any Named Entity.

Abstractive Summarization:

Abstractive summarization process consists of "understanding" the original text and "re-telling" it in fewer words. In abstractive summarization semantic analysis of the document(s) is done on basis of which summary of the document is generated. In this type summarization interpretation of each of the sentence is done and may be represented in the different style from the original one.
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.

2. Literature Review

Byron Georgantopoulos and Stelios Piperidis, "Term-based identification of sentences for text summarization."

This paper methodology described for Greek text which is based on combination of two techniques, terminology extraction and sentence spotting. In this a special kind of abstracts are generated which is a hard NLP task called extracts: set of sentences taken from original context. These sentences are selected on the basis of amount of information they carry about the original context. It uses statically occurrences of terms and several cue phrases which are highlight of sentences to calculate the weight of each sentence and then top scoring sentences are collected and formed the extract.[1]

Gurpreet Singh and Karun Verma, "A Novel Features Based Automated Gurmukhi Text Summarization System."

The research paper is carried out in Punjabi Gurmukhi script. In this paper extraction technique is discussed to summarize the text. In this new features are URL's or email addresses, presence of brackets, inverted commas are included and these are compared with older features. Results are also compared with summary generated by human experts and the overall comparison shown that better in terms of F-score, precision, recall. [2]

Josef Steinberger, Karel Jezek, Using "Latent Semantic Analysis in Text Summarization and Summary Evaluation."

This paper deals with using latent semantic analysis in text summarization. In this paper author describe a generic text summarization method which uses the latent semantic analysis technique to identify semantically important sentences. The proposed method has been further improved. Then author propose two new evaluation methods based on LSA, which measure content similarity between an original document and its summary. In the evaluation part author compare seven summarizers by a classical content-based evaluator and by the two new LSA evaluators. Author also studies an influence of summary length on its quality from the angle of the three mentioned evaluation methods.[3]

Mandeep Kaur and Jagroop Singh, "A survey on different Text Summarized techniques and deadwood is eliminated and remove from the summary."

In this paper, authors propose a system for detection and removal of five different features for the assignment of weight to the sentences. In the next step the highest scoring sentences are selected to form the summary. In the last steps the Deadwood in summaries for Punjabi language. Deadwood means word or phrase that can be omitted without loss in meaning. Removing it shortens and clarifies the summary. Proposed system works in two phases which are: semantic analysis and Adjective Removal Rule.[4]

Ng Choon-Ching & Ali Selamat, "Text Summarization Review."

In this paper author describe 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. [5]

Vishal Gupta and Gurpreet Singh Lehal, "Automatic Punjabi Text Extractive Summarization system."

In this paper author 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 newspaper and fifty Punjabi stories (with 17538 sentences and178400 words). Accuracy of the system is varies from 81% to 92 %.[8]

Zhang Pei-ying and Li Cun-he, "Automatic text summarization based on sentences clustering and extraction."

This paper proposes a solution of text summarization which is a solution to information overload problem. In these sentences clustering based summarization approach is used. It shows that summarization results not only depends upon the sentence features but also on sentence similarity measure This approach consists of three steps: first one is clusters the sentences which is based on semantic distance among sentences in the document and second step is on each cluster calculates the accumulative sentence similarity based on the multi- features combination method and the third step is : it chooses the topic sentences by some extraction rules.[11]

4. Conclusion

In the paper present author conclude that not much of the work has done in the field of text summarization of Hindi language. A system is to be developed that can generate the summary of Punjabi documents on the basis of rules based on Punjabi language. A hybrid technique is need to be developed which uses both extractive as well as abstractive techniques to generate the summary of a Hindi text document that contain information from history. A database is also need to be created which can help the system to assign the weights to the sentences for summary generation.

5. References

[1] Byron Georgantopoulos, Stelios Piperidis ,Term-based identification of sentences for text summarization, Institute for Language and Speech Processing

[2] Gurpreet Singh, Karun Verma, 2014, A Novel Features Based Automated Gurmukhi Text Summarization System, Int. Conf. on Adv. in Comp., Comm., and Inf. Sci. , Elsevier

[3] Josef Steinberger, Karel Jezek, Using Latent Semantic Analysis in Text Summarization and Summary Evaluation, Department of Computer Science and Engineering, Univerzitní 22, CZ-306 14

[4] Mandeep Kaur, Jagroop Singh, A survey on different Text Summarized techniques and deadwood is eliminated and remove from the summary

[5] Ng Choon-Ching , Ali Selamat, Text Summarization Review, Faculty of Computer Science and Information System

[6] Vishal Gupta, Gurpreet Singh Lehal , 2011, Automatic Keywords Extraction for Punjabi Language , International Journal of Computer Science, Issues 8(5) : 327-331.

[7] Vishal Gupta, Gurpreet Singh Lehal ,2011 Named Entity Recognition for Punjabi Language Text Summarization, In International Journal of Computer Applications, 33(3): 28-32.

[8] Vishal Gupta , Gurpreet Singh Lehal, 2012, Automatic Punjabi Text Extractive Summarization system, Proceedings of COLING 2012: Demonstration Papers, pp. 199–206

[9] Vishal Gupta, Gurpreet Singh Lehal,2012,Complete Pre Processing phase of Punjabi Text Extractive Summarization System, International Journal of Emerging Technologies in Web Intelligence

[10] Vishal Gupta, 2013, A Survey of Text Summarizers for Indian Languages and Comparison of their Performance, International Journal of Engineering Trends and Technology (IJETT) – Volume 6 Issue 7.

[11] Zhang Pei-ying, Li Cun-he,2009, Automatic text summarization based on sentences clustering and extraction" IEEE