Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.320 – 323

RESEARCH ARTICLE

Automatic Summarization of Text Documents Written in Hindi Language

Dawinder Kaur[1], Rajbhupinder Kaur[2]

[1]M.Tech Research Scholar, Department of Computer Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

[2]Assistant Professor, Department of Computer Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

[1] dknahar03@gmail.com; [2] er.rajbhupinder@gmail.com

Abstract*— *A summary of a document is a 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 have represented on single-document summaries for the text data written in the Hindi (Devanagri Script). Language contains history information. This paper also presents detection and removal of Deadwood, Extractive and Abstractive summarization methods designed for Hindi language such document contains history information. 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

I. INTRODUCTION

1.1 Summarization

There are a number of scenarios where automatic construction of such summaries is useful. For example, an information retrieval system could present 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 to 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 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. Historical documents in Hindi language contain information which consists of important dates with their associated events "1947, the year in which India got freedom, names of famous persons line "Bhagat Singh" , information that contain names of places like "various battles held in "Panipat" etc. Hence these documents contain a huge amount of information which needs to be summarized so that reader can read and learn the important information from these documents with ease.

Summarization can be two types:
1. Extractive Summarization
2. Abstractive Summarization

I. **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.

II. **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 id 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.

II. LITERATURE SURVEY

Mandeep Kaur and Jagroop Singh "A survey on different Text Summarized techniques and deadwood is eliminated and remove from the summary". In this paper, an author proposes 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. [1]

Visual 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 news paper and fifty Punjabi stories (with 17538 sentences and178400 words). Accuracy of the system is varies from 81% to 92 %. [2]

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. [3]

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. [4]

Vishal Gupta and Gurpreet Singh Lehal," A Survey of Text Summarization Extractive Techniques." 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. [5]

Y. Gong, X. Liu: "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis". In this paper author describe an existing need for text summarizers that small devices 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. 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. [9]

III. METHODOLOGY

The proposed system is based on rule based approach. Handcrafted rules are developed to generate the summary of the text documents written in Hindi language. A corpus for Hindi language is used along with these handcrafted rules to extract the important lines from a text paragraph. The corpus for Hindi language contains the following set of tables:

1. Table contain person names
2. Table contain location names
3. Table contain city names
4. Table contain state and country names
5. Dead phrase along with their replacement words.

To develop rule based approach various formats for names, dates etc. The basic approach generates the summary from a text document written in Hindi language is as follows:
Step 1: Input the Hindi text paragraph from which text summary is to be generated.
Step2: Divide the complete paragraph into lines with the help of end of the line mark (।).
Step3: Remove the dead phrase from the text paragraph.
Step 3: for each line obtained in the step 2 find the equivalent weight for that line by checking that whether it contain named entity, location name, date, city, state or country name and if so then increment the weight of the line by 1 for every entity found.
Step 4: eliminate those lines from the text paragraph which has weight less than that of minimum weight i.e. 5.
Step 5: Combine the other lines as a paragraph and display it to the user.
To implement this methodology a web based interface is used to input the text paragraph from the user. This is implemented with the help of ASP.Net along with c#. To develop the corpus for to store the Hindi named entities MS-ACCESS is used. Proposed system deals with the importance of the line on the basis of the named entities extracted from the text paragraph. More the named entities found in the paragraph the more weight will assign to that particular line. The line which does not contain the minimum weight is not considered in the final result i.e. summary of the paragraph written in the Hindi language.
Secondly dead phrases removal is also implemented to obtain the summary of the text documents written in Hindi language. In this approach dead phrase (combination of two or more words that can be replaced by single word or can be removed) are to be removed from the text paragraph to obtain the more precise results. Dead phrases along with the replacement words are stored in the corpus and hence removed with the help of this corpus from Hindi text paragraph.

IV. RESULTS & DISCUSSION

The proposed system is tested on 60 documents from different domain to evaluate the results. The system removes the 30% - 40% text to obtain the summary of the test. Hindi corpus for named entities contains more the 15000 named entities to generate the summary of the system. The summary text obtained by the system can be also reducing more to 50% by increasing the minimum weight of the lines which in this case set to 5. If this minimum value is set to 7 or 8 the text can be further summarized.

The statistics for proposed system are as follows:

Entity	Numerical Value
Hindi corpus Data Entries	15000+
System testing	60 Documents
Summarized Text	60%-70%
System overall accuracy	91%

The overall system accuracy is achieved to be 91% which is considerably better than that of existing techniques. Online interface to accept the inputs and to provide the outputs is developed. System also displays the total number lines which are input by the user and resultant number of lines which are generated by the system.

V. CONCLUSION & FUTURE SCOPE

The proposed system generates the summary of the text document written in the Hindi language. Rule based approach along with dead phrase removal is used to used generate the summary of the text written in Hindi language. Proposed system gives 91% accurate results when tested on 60 different documents to generate the summary of the Hindi text. Input text size can be reduced to 60% - 70 % with the help of proposed system. System generates the extractive summary of the text given by the user i.e. it does not generate the summary of the text on the basis of the semantics of the text.

As discussed, proposed system generate the summary of the text based only on importance of the data on the basis of named entities extracted in the text paragraph. System does not include the semantic analysis of the Hindi text from which text summary is to be generated. Proposed system generates the summary of the Hindi text obtained from only single document. In future system can be further extended by including the semantic analysis of the text from which the summary is to be generated. System can be improved in such a way that it can generate the summary from multiple documents. Corpus size can be also be further improved which include more dead phrases to generate the more summarized text from the input data. An NER (Named entity Recognition) System for Hindi language can also be integrated with the existing system to improve the overall performance if the system.

REFERENCES

[1] Mandeep Kaur and Jagroop Singh," A survey on different Text Summarized techniques and deadwood is eliminated and remove from the summary."
[2] Vishal Gupta and Gurpreet Singh Lehal," Automatic Punjabi Text Extractive Summarization system." Proceedings of COLING 2012: Demonstration Papers, pages 199–206, COLING 2012, Mumbai, December 2012.
[3] Ng Choon-Ching & Ali Selamat Text Summarization Review
[4]Josef Steinberger, Karel Jezek Using Latent Semantic Analysis in Text Summarization and Summary Evaluation, Department of Computer Science and Engineering, University 22, CZ-306 14
[5] Vishal Gupta and Gurpreet Singh Lehal, (2010). A Survey of Text Summarization Extractive Techniques. In International Journal of Emerging Technologies in Web Intelligence, 2(3): 258268.
[6] Vishal Gupta and Gurpreet Singh Lehal (2011c). Automatic Keywords Extraction for Punjabi Language. International Journal of Computer Science Issues, 8(5) : 327-331.
[7] Vishal Gupta and Gurpreet Singh Lehal (2011d). Named Entity Recognition for Punjabi Language Text Summarization. In International Journal of Computer Applications, 33(3): 28-32.
[8] Vishal Gupta and Gurpreet Singh Lehal (2011e). Feature Selection and Weight Learning for Punjabi Text Summarization. In International Journal of Engineering Trends and Technology, 2(2): 45-48.
[9] Y. Gong, X. Liu: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States 2001, pp. 19-25
[10] R. Radev, S.Teufel, H.Saggion, W. Lam, J. Blitzer, H. Qi, A. Celebi, D. Liu, E. Drabek: Evaluation Challenges in Large-scale Document Summarization. Proceeding of the 41 annual meeting of the Association for Computational Linguistics, Sapporo, Japan 2003, pp. 375-382

