An Improvised Extractive Approach to Hindi Text Summarization

K. Vimal Kumar and Divakar Yadav

Abstract Text summarization is defined as a task of minimizing a text that is produced from one or more texts such that the actual significant information in the texts is not lost. A text summarization tool compresses the text and displays only the important content to the user. Using text summarization, decisions can be made in lesser time and the core of the document be understood. This paper emphasizes on an extractive approach and its implementation on Java. The extractive approach selects the significant sentences based on a thematic approach. Before selecting the thematic words the Hindi stop-words was removed and also the stemming process to retrieve the root words in the sentences under consideration. Stop-word elimination eliminates the semantically null words from the input document and stemming helps in clustering together words with the same radix term. The system is based on an algorithm for scoring the sentences based on occurrence of the radix of thematic words. The sentences with highest score are added to the summary. The generated summary is further processed based on removal of extraneous phrases from the previously selected summary sentences so as to bring the sentences closer to human generated summary. The testing of the accuracy of the system can be made by using a technique called The Expert Game. In expert game, experts underline and extract the most interesting or informative fragments of the text. The recall and precision of the system's summary is measured against the human's extract. Based on the testing, the system is found to be 85 % accurate.

Keywords Hindi text summarization · Extractive approach · Thematic approach

K. Vimal Kumar (✉) · D. Yadav
Jaypee Institute of Information Technology, Noida, India
e-mail: vimalkumar.k@gmail.com

D. Yadav
e-mail: divakar.yadav@jiit.ac.in

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_28

1 Introduction

There is need of efficient automatic text summarization as the internet provides the access to a very large amount of data in a particular language. People in today's world invest based on stock market updates and they go to movies, various tourist places on the basis of reviews they've seen. This type of text summarization tool helps them in making decisions in a lesser duration. There is an extensive amount of information available in Hindi as well. There is a growing need to make important decisions in time constrained environments and also understand the gist of a Hindi document in given time without missing out any important information. There are very few summarization systems that target Hindi language. This paper targets the problem of information overload and proposes a system for extractive text summarization which compresses the input document but not losing the important content in the document. As access to data from anywhere has increased so the demand for an automatic text summarization has also increased. Automatic summarization system reduces a text document or a multiple documents into a short set of texts or paragraph that conveys the actual semantics of the text and also should not lose the main meaning described in the text. The compressed text must be able to convey the meaning contained in the original text and not be difficult to comprehend.

Automatic text summarization is the process of shortening a given text, by a computer program, without losing the actual information to be conveyed. There are two widely used methods in text summarization—Extractive and Abstractive. Extractive summarization extracts the texts and creates the summaries by reusing portions (words, sentences, etc.) of the input text, while abstractive summarization is those which create the summaries by re-generating the significant content of the input text. In case of summarization system, there are summarises from single document or multiple documents and these kind of summarization system is called as multi document summarization system.

Majority of the research work carried out so far has been emphasized more on widely used English and other European languages. Indian Languages have been explored little because of the amount of information available in non-English language were less. However, the scenario is now changing and a large amount of information has become available in various languages. The need for text summarization methods that can handle Indian languages appear to be growing. We aim to develop a system for automatic summarization based on extractive summarization techniques for creation of summary of Hindi text documents so that the user can view the summarized form rather than in full. The system would currently produce summary for single text documents and in Hindi. In case of decision making systems based on analysis of large text, Hindi text summarization plays a vital role for the analysis process. Apart from this, Hindi text summarization has various applications in those systems where there is requirement for text analysis and knowledge representation. This system is based on extractive summarization and it attempts to identify the set of significant sentences that are most important for

the understanding of a given document in Hindi. In case of extractive summarization, the identification of significant sentences plays a major role in improving its accuracy.

2 Related Work

A lot of summarization approaches exists, but mostly for English and European language. Sentence location based approach works for news articles [1]. Location heuristics used in such a method is: Considering the newswire articles, the first sentence is often taken as the most important sentence and in case of technical articles, last few sentences in the abstract or those from conclusions provide the main details contained in the document. Sentences that are relevant to the title of the text will be extracted from the text and such an approach is called as title keyword approach [2, 3]. Apart from title of the text, the upper case words which contain the acronyms and proper nouns can also be used as a parameter. But such criteria won't work for Indian languages as the proper noun or acronym can't be identified by the upper case letter. In multi document summarization, the system clusters the similar documents and then extracts the important sentences from these clustered documents to form its summary. Such a method is known as Cluster based method. In query based text summarization system, the sentences are scored based on the frequency count of terms (words or phrases). In this type of system, the sentences containing the query phrases are given higher scores than those containing single query words. Based on the sentence scores, the output summary is formed using the high scored sentences.

Universal networking language based approach is basically used for language independent system [4]. This system generates a multi lingual summary by using an Interlingua document representation language called "Universal Networking Language" (UNL). In a graph theoretic approach, each sentence is denoted as a node and two such nodes are connected with an edge if the two sentences corresponding to those nodes share some common words, or similarity. But in all these methods the identification of sentences that has to be extracted is very important. The important sentences or documents can be identified using TF-ISF (Term frequency inverse sentence frequency) method and this method has been adapted from the information retrieval idea of calculating TF-IDF [5].

Since this system focuses on Hindi language, there is a need to bring together different lexical and semantic relations between various words in a system and that system is Hindi WordNet. WordNet organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English WordNet. Hindi WordNet was explored to understand the usage of Hindi WordNet API to make use of the dataset and available synset, hypernyms, hyponyms etc.

3 Proposed Summarization System

Summarization is of two types, abstractive summarization and extractive summarization. Extractive summarization works based on selection of a subset from the existing words, phrases, or sentences in the original text to form the summary. In contrast, an abstractive summarization uses natural language generation techniques to create a summary of a text that is closer to what a human might generate. The task of extractive summarization generates the text summary by concatenating various extracted subset of text segments from the input text [6]. The selection of sentences would be based on single words or multiword expressions. The proposed solution to the summarization system is a thematic term based approach which is based on frequent term based approach [7] for summarization of single Hindi text document. The system is divided into four major steps: pre-processing, thematic word generation, sentence scoring and summary generation. The system architecture for the proposed algorithm is as shown in Fig. 1.

Pre-processing is performed in two steps. First step is stop-word elimination in which semantically null words in Hindi are removed. A list of 170 stop words was used to perform stop-word elimination. The text document after removal of stop-words is input for stemming, in which all the remaining words are converted into their morpheme term. This is performed using longest suffix stripping method [8, 9]. The document thus obtained contains only the radix terms which is used to determine the thematic terms of the document. This is done by recording the occurrence of each term in the document making use of a document-term matrix. The chosen words and their synset are used to score the sentences.

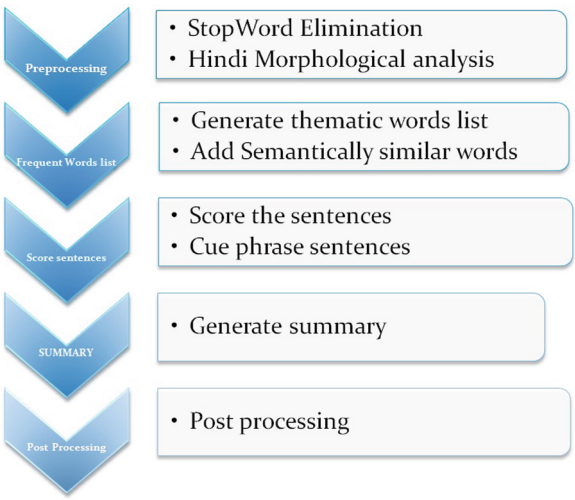

Fig. 1 System architecture

Scoring the sentences is done using the equation given in [1] which is as follows:

$$Sc_j = \sum M[i, j]/|Terms| \quad (1)$$

where,

Scj	Score of sentence j
M[i, j]	value of the cell [i, j]
|Terms|	total number of terms in the document

The top scored words are determined based on the threshold, input by the user. There are sentences which contribute more in the understanding of the document and such sentences have cue phrases like “nishkarsh”, “mahatvapoorna”, “natija”, “parinaam swarup”, “udahran”. These kinds of cue phrases are also included in the summarized version as these sentences contain conclusive remarks and examples [10]. The extracted sentences are output based on their relevance in the input document. Thus the summary is generated for the input document.

The major factor that decides about the efficiency of summarization system is the compression ratio. To achieve the desired compression ratio, the output sentences are further sent for post processing using the local context information [11, 12]. The connector (clause) including connecting words like: कि, जब- तब, अब- तब, तो, पर phrases are identified to select the unimportant text in the summary generated. To reconfirm the selection, the sentence score is also checked for these unimportant clauses and thus would be removed if the score found is less. Also, the words in the extracted sentences are linked with words in its local context. But this may lead to identification of those sentences which are semantically repeated and morphologically related. These sentences are linked through one of the lexical relations based on the morphological relation between sentences.

During this post processing phase, the extraneous phrases are identified on the basis of clauses and connecting words such as “ki” “par” “kintu” “magar” “tab” etc. The phrases are scored based on the relevance to the theme of the main document. The relevance of each word in the phrase is found with the local context and scored based on its relation to the thematic words. An example of removal of extraneous phrases is given below:

Original sentence

हैरानी इस बात की है कि अंतरराष्ट्रीय क्रिकेट मैचों के दौरान हृदय में देशभक्ति का भाव लेकर देखने वाले लोग अब ऐसी स्थानीय स्तर के प्रतियोगिता भी मनोरंजन के लिये देखने लगे हैं

Reduced sentence

अंतरराष्ट्रीय क्रिकेट मैचों के दौरान हृदय में देशभक्ति का भाव लेकर देखने वाले लोग अब ऐसी स्थानीय स्तर के प्रतियोगिता भी मनोरंजन के लिये देखने लगे हैं

This process of removing extraneous phrases is done using features of Hindi WorldNet. The identified extraneous phrases are then scored and the phrases having minimum score are removed from the summarized output. By considering the number of links between the words and the type of relation between various words, the system computes the score for each word in the extracted sentence based on the formula provided by Jing [13]:

$$\text{Context Weight}(w) = \sum (L_i \times \text{NUM}_i(w)) \quad (2)$$

where,

i	total number of lexical relations types identified
Li	Weights assigned for various types of lexical relation
NUMi(w)	Number of particular type of lexical relation between w and various words in the input sentence

Extraneous phrases are scored on the basis of following linkages:

- Inflectional Relation to thematic words
- Synset: It is a set of synonymous words. For example, "वद्यिालय, पाठशाला, स्कूल"
- Hyponymy and Hyponymy: बेलपत्र is a kind of पत्ता means पत्ता is a hypernym and बेलपत्र is the hyponym.
- Meronymy and Holonymy (Part-whole relation): जड़ (root) is the part of पेड़ (tree), meaning that जड़ (root) is the meronym of पेड़ (tree) and पेड़ (tree) is the holonym of जड़ (root).
- Antonymy: Antonymy is a relation that holds between two words that (in a given context) express opposite meanings

If a phrase or word is found to be strongly related to the local context, it is not removed from the summary sentence. The final summary is thus generated after post-processing.

4 Implementation Details

For implementing the above discussed approach, Java version 1.6.18 as programming language and Netbeans 6.9 as platform was used as they supports UTF-8 format, which is necessary for Processing Hindi text. Further Hindi WordNet API is used to determine lexical linkages of the words and phrases with respect to the local context mentioned in the input text and also to determine hypernyms, hyponyms etc.

Screenshots of implementation are given in figures below. In the upper left side window as shown in Fig. 2, user chooses the text document to be summarized.

In the middle left side window the summary is presented to the user along with the identified keywords highlighted in the bottom left side window as shown in Fig. 3.

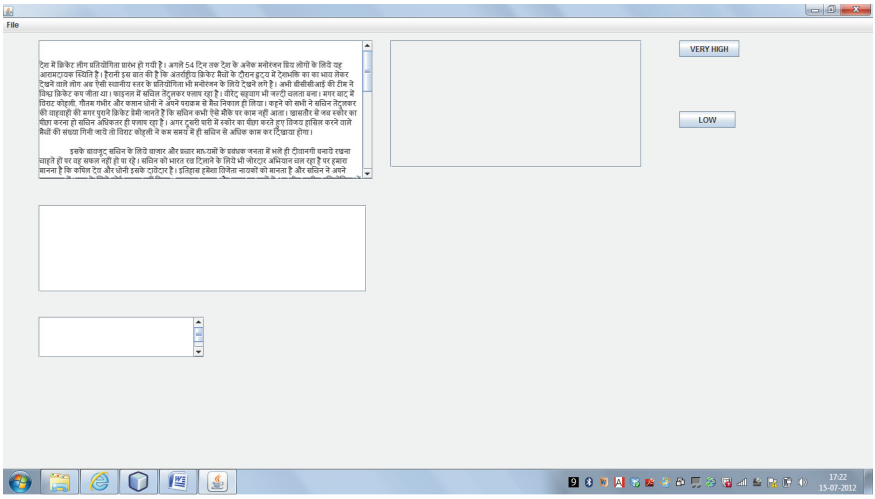

Fig. 2 Test case1-input

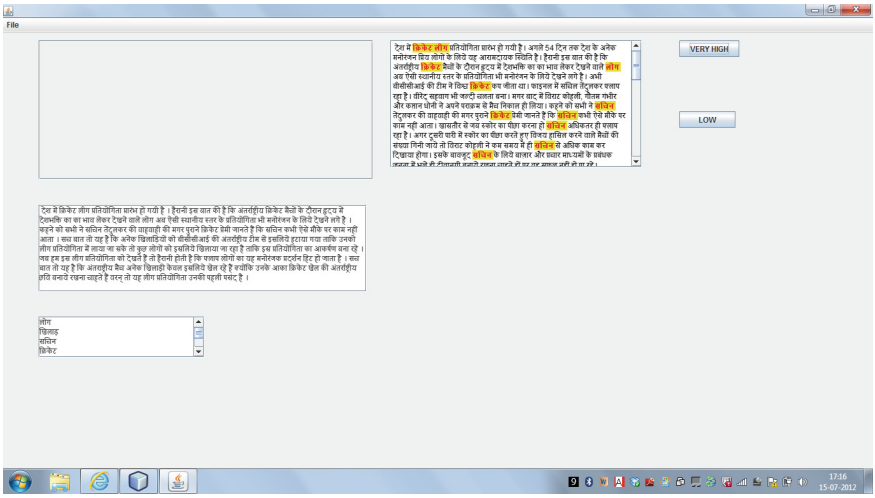

Fig. 3 Test case1-output (accuracy: 85.64 %)

Once the input document is chosen, the stop words are eliminated from the document. For the stop word elimination, a list of all the stop words for Hindi language is generated which includes semantically null words. The stop words list consists of 170 words. The list includes sample words like “par”, “inhey”, “jinhey”, “ke”, “pe”, “yeh”, “hain”, “veh”, “ityadi”, “dwara” etc. These words were eliminated before generating frequent words list. Since the stop words and input text

document is in Hindi the entire processing needs to be done in UTF-8 format. The resultant words generated are the semantically important words.

For a sample input sentence "राज्यसभा सचिवालय में नामांकन पत्र की जांच का काम अब मंगलवार तक के लिए टाल दिया गया है।", the words generated after stop word elimination are "राज्यसभा", "सचिवालय", "नामांकन", "पत्र", "जांच", "काम", "अब", "मंगलवार", "टाल".

The second step of pre-processing which is Hindi morphological analysis is performed using the native method longest suffix matching algorithm. Suffixes are found by clustering the suffixes related to a category. Nouns are inflected based on the case, the number, and the gender. Suffixes are identified to reducing the inflected forms of masculine nouns, feminine nouns, adjective inflections, and verb inflections to a common root.

The root words obtained are the basic keywords and for these words the frequency of occurrence is calculated so as to determine the words which occurs the most in the document. The top most frequent terms generated are then chosen and their semantically similar words are also added to the frequent terms list. The top scored sentences are generated as part of the summary in the order of presence in the original text document.

For evaluation of the summary generated a human analysis technique was used. A gold standard summary was generated by humans and the summary of the system was compared with the system generated summary. The number of lines matching and the total number of lines were used to evaluate the system summary.

During post processing, the tagging system is used to identify the possible adjectives, nouns, verbs and adverbs so that sentences can further be trimmed for a better summary [14]. The score for each phrase is calculated by adding up the score of each word in the phrase. This score indicates how important the phrase is in local context. Based on this score the unimportant phrases can be removed from the extracted sentence.

5 Result Evaluation

The generated summary was tested for its accuracy by comparing it with a gold standard summary line by line. The accuracy of the system is tested using an approach called The Expert Game. Ask experts to underline and extract the most interesting or informative fragments of the text. Measure recall and precision of the system's summary against the human's extract. The system is found have 85 % match between gold standard summary and system generated summary.

As shown in Table 1, the average accuracy of the system with the expert's manual summary is found to be 85 % and also the average retention ratio is 81.1 %. This is found to considerably good percentage when compared with the system without the post processing stage. In case of post processing stage the system combines the valuable information by maintaining the same desired compression ratio.

Table 1 Analysis of system accuracy for various input documents

Document ID	Retention ratio (%)	Accuracy (%)
1	81.08	85.64
2	78.94	85.71
3	83.33	85.15
4	81.08	83.33

6 Future Work

The future plan is to develop the multi-document summarization system using this proposed algorithm. For which the extracted summary has to be generated from multiple documents those are related to the context. The enhancement of the system to multiple languages is also a part of the future scope of work involved. The generated summary could be in multiple languages and the system could work on multiple languages to generate summary [15, 16]. This system in combination with a language translation system can be used to generate the summary in a language of the user's choice independent of the source language document. It could involve web mining also to extract the summaries from multiple documents available online.

References

1. Lloret, E., Palomar, M.: Finding the best approach for multi-lingual text summarisation: a comparative analysis. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria (2011)
2. Lloret, E., Palomar, M.: Text summarisation in progress: a literature review. Artif. Intell. Rev. **37**(1), pp. 1–41 (2012). ISSN: 0269-2821
3. Alguliev, R.M., Aliguliyev, R.M.: Effective summarization method of text documents. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), pp. 1–8 (2005)
4. Mangairkarasi, S., Gunasundari, S.: Semantic based text summarization using universal networking language. Int. J. Appl. Inf. Syst. **3**(8), 18–23 (2012) (Published by Foundation of Computer Science, New York, USA, August 2012)
5. Juneja, V., Germesin, S., Kleinbauer, T.: A learning-based sampling approach to extractive summarization. In: Proceedings of the NAACL HLT 2010 Student Research Workshop, pp. 34–39 (2010)
6. Gupta, V., Lehal, G.S.: Survey of text summarization extractive techniques. J. Emerg. Technol. Web Intell. **2**(3), pp. 258–268 (2010)
7. Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.: Summarization text documents: sentence selection and evaluation metrics. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), Berkeley, USA, 15–19 Aug 1999, pp. 121–128
8. Ramanathan, A., Rao, D.D.: A lightweight stemmer for Hindi. In: Proceedings of EACL (2003)
9. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)

10. Gupta, V., Lehal, G.S.: Features selection and weight learning for Punjabi text summarization. Int. J. Eng. Trends Technol. **2**(2), 45–48 (2011)
11. Chen, F., Han, K., Chen, G.: An approach to sentence selection based text summarization. In: Proceedings of IEEE TENCON02, pp. 489–493 (2002)
12. Jing, H.: Sentence reduction for automatic text summarization. In: Proceedings of the 6th Applied Natural Language Processing Conference (2000)
13. Jing, H.: Cut-and-paste text summarization. Ph.D. thesis, Department of Computer Science, Columbia University, New York (2001)
14. Ray, P.R., Harish, V., Basu, A., Sarkar, S.: Part of speech tagging and local word grouping techniques for natural language processing. ICON (2003)
15. Patel, A., Siddiqui, T., Tiwary, U.S.: A language independent approach to multilingual text summarization. Conference RIAO2007, Pittsburgh, PA, USA (2007)
16. Mihalcea, R., Tarau, P.: An algorithm for language independent single and multiple document summarization. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Korea (2005)