Text Summarization of Hindi Documents using Rule Based Approach

Manisha Gupta
Department of Computer Science
GZSCCET,
Bathinda, Punjab, India
manishagupta348@gmail.com

Dr.Naresh Kumar Garg
Department of Computer Science
GZSCCET,
Bathinda, Punjab, India
naresh2834@rediffmail.com

***Abstract*—Automatic summarization plays an important role in document processing system and information retrieval system. Generation of summary of a text document is a very important part of NLP. There are a number of scenarios where automatic construction of such summaries is useful. Text summarization is that process which convert a larger text into its shorter form maintaining its information. Summary of a longer text saves the reading time as it contain lesser number of lines but all important information of the original text document. In this paper we present a novel approach for text summarization of Hindi text document based on some linguistic rules. Dead wood words and phrases are also removed from the original document to generate the lesser number of words from the original text. Proposed system is tested on various Hindi inputs and accuracy of the system in form of number of lines extracted from original text containing important information of the original text document**

Keywords—Text Summarization; NLP; Rule based approach; Dead Wood Removal; Dead Phrase Removal

I. INTRODUCTION

A summary of a document is a (much) shorter text that conveys the most important information from the source document. There are a number of outlines where automatic construction of such summaries is useful. For example, system of an information retrieval could present an automatically built summary in its list of retrieval outcome, for this user to firstly decide which documents are interesting and worth opening for a closer look—that is what Google models to some degree with the brief extracts shown in its search results. Other examples include automatic creation of summaries for news articles or email messages to be sent to mobile devices as SMS; brief extracts of information for businessmen, government officials, researches, etc., and summarization of pages of web to be shown on the screen of a mobile device, and many others.

Text summarization tasks can be categorized into single-document and multi-document summarization. In single-document summarization, the summary of single document is to be built, while in multi-document 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 thesis we have experimented with single-document summaries for the text data written in the Hindi Language.

Methods of summarization can be classified into abstractive and extractive summarization. An abstractive summary is based on random choice text that describes the contexts of the source document. Abstractive summarization process comprises of "perceiving" the real text and "re-telling" it in fewer words. Namely, an abstractive summarization method uses semantic methods to examine and interpret the text and then to find new concepts and expressions to illustrate it by generating a new brief text that conveys the most important information from the original document. While this may seem the right way to construct a summary (and this is how human beings do it), in real-life setting immaturity of the corresponding semantic technology for text analysis and generation currently renders such methods practically infeasible.

An extractive summary, on other hand, is a selection of sentences (or phrases, paragraphs, etc.) from the original text, generally presented to the user in the same order—i.e., as that of the source text with most sentences omitted. An extractive summarization method, for each sentence, whether it will be included in the summary or not . The resulting summary reads rather awkward; however, simplicity of the underlying statistical techniques makes extractive summarization an good, well made, language-independent alternative to more "intelligent" abstractive methods. In this thesis work, we consider extractive summarization for the text documents.

A normal extractive summarization method consists in various ways, at each of them different options can be chosen. We will presume that the units of selection are sentences (it could be phrases or paragraphs). Thus final aim of the extractive summarization process is selection of sentence. One way is to select the appropriate sentences is to allocate some numerical measure of value of a sentence for the summary and then select the appropriate ones; the process of allocating these weights to the word of a sentence is called sentence weighting. Another one is to estimate the usefulness of a sentence is to sum up usefulness weights of discrete terms of which the sentence consists; the process of estimating the single terms is called term weighting. For this, one should decide what are the terms: for example, they can be phrase,words; deciding what entity will count as terms in the task of term selection. Different extractive summarization methods can be characterized by how they perform these tasks.

978-1-5090-3411-6/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMETE.2016.104

II. LITERATURE SURVEY

Gurmeet Singhand Karun Verma (2014) represented the new features in processing phase in Punjabi text extractive system which have two phases 1) Pre Processing 2) Processing. In this paper term preprocessing is defined as the phase which recognize the word boundary, sentence boundary, Punjabi stop words elimination and root word identification and in the second phase i.e. processing phase, adds the new feature to which they are calculated and a weight is assigned to each sentence on the authority of which unwanted sentences are eliminated from the input text. It has been tested over twenty Gurumukhi Documents comprises of news (achieves high rates over 90%) and for stories (achieves range above 94%) and for articles (the accuracy achieved was 82.04%,) taken from Internet.

MandeepKaur and Jagroop Singh (2013) propose a system in which first it will provide the summary by applying five different features. The weight is calculated for particular sentence and summary is generated. After that the deadwood rules applied on the generated summary to eliminate the deadwood words and phrases in the Punjabi text document. Deadwood means that the word or phrase which has no meaning. By eliminating these words and phrases it will shorten the text length but the meaning remains same.

Vishal Gupta and Gurpreet Singh Lehal (2013) concentrate on single document multi news Punjabi extractive summarizer. Very first time system is developed. For the first time new resources developed like Punjabi noun morph, Punjabi stemmer and Punjabi named Punjabi keywords identification entity recognition, , normalization of Punjabi nouns etc. Any Punjabi newspaper contains hundreds of multiple news of different length. Different compression ratio is selected and starts with the extraction of headlines of news and lines just next to headlines System defines into two main steps: Pre Processing and processing phase. Pre Processing phase shows the Punjabi text in structured way. Statistical features and linguistic features are relative sentence length feature, Punjabi keywords identification, numbered data feature and Punjabi headlines identification, identification of lines just next to headlines, identification of Punjabi-nouns, identification of Punjabi-proper-nouns, identification of common-English-Punjabi-nouns, identification of Punjabi-cue-phrases and identification of title-keywords respectively. Scores are being calculated by using sentence weight feature like mathematical regression. High scored sentences are selected for the final summary. In final summary, sentence locations are maintained. The analysis of Punjabi corpus, Punjabi dictionary and Punjabi noun-morph for developing these resources. These Punjabi resources have been developed for the first time and these might be helpful for developing other NLP applications for Punjabi language.

Vishal Gupta and Gurpreet Singh Lehal (2012) have given the approach to new features in Punjabi Text Summarization for its growth.As already some of the features being used to extract the summary. For the very first time these features are being added in this author proposes a system with these features to get best result in the summary and these features are based statistical and linguistic. Pre Processing is structured way to present the original Punjabi text document. It includes Punjabi sentences boundary identification, Punjabi words boundary identification, Punjabi stop words elimination, Punjabi language stemmer for nouns and proper names, applying input restrictions and elimination of duplicate sentences. Many of these are to be done from the scratch as no work is done in this yet. This is first time that these resources which have been developed for Punjabi and these can be beneficial for developing other Natural language processing applications for Punjabi.

III. PROPOSED METHODOLOGY

The proposed system is based on rule based approach. Handcrafted rules are developed to create the summary of the text documents written in Hindi language. A corpus for Hindi language is used in addition with these camp-made 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.
6. Tables Contain Birds/Animals Name
7. Table contain special Symbols
8. Table contain Momentary Expressions
9. Table Contain measurement values.

In every paragraph all sentence is assigned a weight according to certain features. These features are attributes that attempt to represent the data used for their task. We discover five features for each sentence. Each feature is given a value or score. These five features are as follows:

A. Identification of Hindi Titleword (S1)

In this feature all the sentences in the Hindi paragraph are identified which contains the title word. Then scores are identified for those sentences. The words which occur in the sentences also in the title give high scores. The score for this feature can be calculated as the ratio of the unique number of words in the sentence that occur in the title over the number of words in title.

$$\text{Score (S1)} = \frac{\text{No. of title words in the sentence}}{\text{No. of words in the title}}$$

B. Identification of the number of words in the Sentence (S2)

The score for this feature can be calculated as the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

$$\text{Score (S2)} = \frac{\text{No. of words in Sentence}}{\text{No. of words in largest sentence}}$$

C. Cue words (S3)

The cue words are the index words which occur in the sentence increases the importance of the sentence. Cue words are नतीजा, नतीजे, निचौड़ and अनुबंध etc

$$\text{Score (S3)} = \frac{\text{No. of Cue Words in the sentence}}{\text{Length of the sentence}}$$

D. Common English Hindi Noun Feature (S4)

English words are now commonly being used in Hindi. For example consider a Hindi language sentence such as टैकनालोजीकेयुगमेंमोबाईल (Technology ke yug mein mobile). This sentence contains टैकनालोजी(Technology) and मोबाईल (mobile) as English-Hindi nouns. Also these should obviously not be coming in Hindi dictionary. Common English-Hindi noun words are helpful in deciding sentence importance. Common English-Hindi noun feature score is calculated by dividing number of common English-Hindi nouns in a sentence with length of that sentence.

$$\text{Score (S4)} = \frac{\text{Common English Hindi nouns in Sentence}}{\text{Sentence length}}$$

E. Position (S5)

In this feature the place of the sentence is recognized whether it is located at the starting or end or middle of the paragraph. The sentences at the starting and the end of the paragraph are given highest scores. If there are 7 sentences in the paragraph then the sentence score can be calculated as:

Score (S5) for 1st sentence - 7/7
Score (S5) for 2nd sentence -6/7
Score (S5) for 3rd sentence -5/7
Score for (S5) 4th sentence -4/7
Score for (S5) 5rd sentence-5/7
Score for (S5) 6th sentence-6/7
Score for (S5) 7th sentence-7/7

F. Numeric Data (S6)

In this feature sentences containing the numeric data are identified. Sentences containing numeric data are important and have higher probability to be extracted for the summary. E.g. दिल्ली में 7 फरवरी को विधान सभा की चुनाव में70 सीटो में आपको 40 सीटे भाजपा को 25 औरकांग्रेस को 5 सीटे आने का अनुमान है ।In this sentence 7, 70, 40, 25 and 5 are numbers.

The score for this feature can be calculated as the ratio the number of numerical data that occur in sentence over the sentence length.

$$\text{Score (S6)} = \frac{\text{No. of Numerical data in sentence}}{\text{Sentence Length}}$$

G. Nouns in Sentence (S7)

A noun is the name of the person, place or thing. Sentences having nouns are important and have higher probability to be extracted for the summary. Nouns are मनवीर, सच्चेनदरा, यौगी, सनदीप, मयूर, शिवांगी, चारू etc.

The score for this feature can be calculated as the ratio the number of Noun words that occur in sentence over the sentence length.

$$\text{Score (S7)} = \frac{\text{No. of Noun words in sentence}}{\text{Sentence Length}}$$

H. Presence of Brackets (S8)

Sentences sometimes may contain brackets such as ( ) parentheses, {} curly brackets etc. mostly braces contains material which could be omitted without destroying or altering sentence meaning. After doing analysis it has been found that brackets do not contain important information and has lower probability to be included for the summary. The feature score can be calculated as the ratio sentence length minus the total no. Of words within bracket that occur in sentence over the sentence length.

$$\text{Score (S8)} = \frac{\text{Sentence Length - Total no of words within brackets in sentence}}{\text{Sentence Length}}$$

I. Presence of Inverted Commas (S9)

In Hindi (“ ”,‘ ’)quotation marks or inverted comma surrounding quotations, literal title ,direct speech, or name etc. contains important information. After doing study it has been seen that an inverted comma has higher probability to be included for the summary. The score for this feature can be calculated as the ratio total number of words within quotation that occur in sentence over the sentence length.

$$\text{Score (S9)} = \frac{\text{No. of words in quotation marks or inverted commas}}{\text{Sentence Length}}$$

J. Keywords in Sentence (S10)

Keywords are words that appear with unusual frequency (very high) in a text document. Keywords identification and calculation is very important feature and it helps in deciding sentence importance. In proposed system top 10% words having higher frequencies are taken as key-words.

$$\text{Score (S10)} = \frac{\text{No. of Keywords in the sentence}}{\text{Total number of words in the sentence}}$$

K. Similarity between Two Sentences (S11)

Similarity between two sentences is used to determine whether semantically these two sentences are equal or not. Root words are used for determining similarity between two Sentences. If two Sentences having utmost root words match then they have higher probability of being similar.

L. Presence of URL’s or Email Addresses (S12)

Internet is important and widely used application now days. Text document may have URL’s or Email Addresses present in it, which provides more information about the document in process. After doing analysis of various Hindi newspapers and Hindi documents it has been found that this feature has very high importance than other and needs to be extracted for the summary.

In this case system will increase the weight of sentence by 1 for each URL or email address found in it.

M. Animal/Bird Noun (S13)

A noun may be the name of the animal and bird. Sentences containing nouns are important and have higher probability to be extracted for the summary. In Hindi nouns are चिड़िया, शेर, चिता, हाथी, मोर etc. In this case system will increase the weight of sentence by 1 for each Animal/Bird Noun found in it.

$$\text{Score (S13)} = \frac{\text{No. of Noun words in sentence}}{\text{Sentence Length}}$$

N. Use of Special symbol (S14)

In this feature special symbols like $, %, & etc. been introduced and After doing analysis it has been found that an special symbols has higher probability to be included for the summary. The score for this feature can be calculated as the ratio total number of symbols that occur in sentence over the sentence length.

$$\text{Score (S14)} = \frac{\text{No. of Special Symbol in sentence}}{\text{Sentence Length}}$$

Sentence Length

O. Measurement Values (S15)

In this feature sentences containing the values like 50 km has high probability to introduce in the summary. The score for this feature can be calculated as the ratio total measurement values that occur in sentence over the sentence length. Eg. 50 किमी and 10 लीटर etc

$$\text{Score (S15)} = \frac{\text{Total no. of measurement values in sentence}}{\text{Sentence Length}}$$

P. Monetary Expression (S16)

In this feature sentences containing the values like एक रूपया,सौ डालर,पचास रूपये has high probability to introduce in the summary. The score for this feature can be calculated as the ratio monetary expression that occurs in sentence over the sentence length.eg. एक रूपया,सौ डालर,पचास रूपये

$$\text{Score (S16)} = \frac{\text{Total number of monetary expression in sentence}}{\text{Sentence Length}}$$

Q. Date Format (S17)

Dates are very important for any historical documents. Presence of dates in the sentence increases the importance of the sentence. This feature is considered as an important feature for summarization of text. Rules for various types of date formats are to be developed to recognize the various date's patterns.

$$\text{Score (S17)} = \frac{\text{Total number of dates in sentence}}{\text{Sentence Length}}$$

R. Conjunction (S18)

Conjunction is referred to as the combination of more than one sentence into a single unit. It is considered that if conjunction keywords are present in the sentence it is considered as an important sentence.

$$\text{Score (S18)} = \frac{\text{Total number of conjunction keywords in sentence}}{\text{Sentence Length}}$$

Sentence-Extraction Phase

In Sentence-Extraction phase firstly final weight of every sentence is calculated using Weight-Ranking equation. After calculating final weight of every sentence, extraction of sentences is done according to compression ratio required.

Sentence weight (Si)= S1+S2+S3+S4+S5+S6…….. + S20

Where Sentence Weight (Si) is a final weight of sentences (Si) and f1,f2.... f18 are features which are computed above.

Selecting best sentences to include in the summary

The number of sentences to be included in the summary can be calculated as the ratio of the number of sentences in the Hindi paragraph over the amount depends on the percentage selected by the user. The user can select 25%, 33%, 50% sentences to be included in the summary of the original text.

Sentences to be included in the summary are those sentences which have highest scores. The number of sentences to be included in the summary can be calculated as ratio of number of sentences in the Hindi paragraph over 3.

Summary length =Total No. of sentences in Paragraph 3

Summary is 1/3 part of the source paragraph. It contains the 33% data from the source Hindi paragraph to convey the meaning of the paragraph.

IV. Results and Discussion

The proposed system is tested on 30 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 reduce 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	30 Documents
Summarized Text	60%-70%
System overall accuracy	96%

The overall system accuracy is achieved to be 96% 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 of lines that can entered by the user and resultant number of lines which are generated by the system.

V. Conclusion and Future Scope

A. CONCLUSION

The proposed system generates the summary of the text document written in the Hindi language. Rule based approach with dead phrase and deadwood removal is used to used generate the summary of the text written in Hindi language. Proposed system gives 96% accurate results when tested on 30 different documents to generate the summary of the Hindi text. Input text size can be decreased to 60% - 70 % with the help of proposed system. System generates the extractive summary given by the user i.e. it does not generate the summary of the text on the basis of the semantics of the text.

B. FUTURE SCOPE

As discussed, proposed system generate the summary of the only on the basis of named entities extracted in the text paragraph. Proposed System does not have the semantic analysis of the Hindi text from which 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 expanded 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 upgraded which include more dead phrases to generate the more summarized text from the input data. A Named entity Recognition (NER) System for Hindi language can also be concatenated with the existing system to enhanced the overall performance of the system.

References

[1] Singh, S. and Shan, H. S. (2002) *"Development of Magneto Abrasive Flow Machining Process",* International Journal of Machine Tools & Manufacturing, vol. 42, issue 2, pp. 953-959.

[2] Laroiya, S.C. and Adithan, M. (1994), *"Precision Machining of Advanced Ceremics"* Proceeding of the International Conference on Advanced Manufacturing Technology (ICMAT - 94), University Teknoloi Malaysia, Johor Bahru ,Malaysia, pp 203-210.

[3] Adithan, M. and Gupta, A.B. (1996), *"Manufacturing Technology",* New Age, International Publishers, New Delhi

[4] Gurmeet Singh and Karun Verma (2014), "*A Novel Features Based Automated Gurmukhi Text Summarization System",*International Conference on Advance in Computing, Communication and Information Science, Elsevier, 2014 pp 424–432.

[5] MandeepKaur and Jagroop Singh (2013),"Deadwood Detection and Elimination in TextSummarization for Punjabi Language", International Journal of Engineering Sciences, Vol. 8, Issue June 2013.

[6] Vishal Gupta and Gurpreet Singh Lehal (2013), "*Automatic Text Summarization System forPunjabi Language",* Journal Of Emerging Technologies In Web Intelligence, VOL. 5, NO. 3.

[7] Vishal Gupta and Gurpreet Singh Lehal (2012), "*Automatic Punjabi Text Extractive Summarization System",* Proceedings of COLING 2012, Mumbai, December 2012. pp 191–198.

[8] ChetanaThaokar and Latesh Malik, "Test model for summarizing hindi text using extraction method", In Information & Communication Technologies (ICT), 2013 IEEE Conference on, pp 1138–1143. IEEE, 2013.

[9] Vishal Gupta and Gurpreet Singh Lehal (2012), "Complete Pre Processing phase of Punjabi Text Extractive Summarization System", Proceedings of COLING 2012, Mumbai,pp 199–206.