Hindi Multi-document Text Summarization using Text Rank Algorithm

Aniket Suryavanshi
Department of Information Technology
Vidyalankar Institute of Technology,
Wadala, Dadar East.

Bhavika Gujare
Department of Information Technology
Vidyalankar Institute of Technology,
Wadala, Dadar East.

Allan Mascarenhas
Department of Information Technology
Vidyalankar Institute of Technology,
Wadala, Dadar East.

Bhanu Tekwani
Department of Information Technology
Vidyalankar Institute of Technology,
Wadala, Dadar East.

ABSTRACT
Summarization is a technique in which the original text is summarized by selecting only what is important from the text. In today's world with the great advancement in the field of technology, huge amount of information is available online. In order to get only the necessary information from one or more sources available online, summarization is needed. By taking into the consideration the above problem, technique for hindi text summarization is proposed. Various documents related to sports, essays, news etc. which are available online were taken input for the system. The extraction of the summarized text was done on the basis of some important features of the text. It can be further used for news summary, summaries of books, etc.

Keywords
Hindi text documents; text rank algorithm; Extractive method; multi document

1. INTRODUCTION
When the Internet was introduced, the world was opened to data sources in an expanded way where huge amount of information is easily made available on the internet. People can extract variety of information from the internet ranging topics like Politics, Bollywood, Sports, Current Affairs etc. There are multiple sources like websites and app that provides information about the current affairs which can be seen on the internet for multiple days. But now the times have changed. In this growing competitive and technical world people tend to focus on highlights and expect the long story short due to the time constraints. This brings to the arrival of highly automated world. Many models have been developed to automatically summarize the English Language Data although the challenge is to summarize the Hindi language text. Here, a model has been proposed for the same which can automatically summarize the Hindi Text. Text summarization can be achieved by two methods: a) extraction based i.e. pulling key phrases from the text and combining them to make a summary. But this process sometimes give grammatical errors as it only considers the key phrases and simply combines them together.b) abstraction based i.e. this entrails paraphrasing and shortening the parts of the document using natural language processing concepts and methods to generate more generalized text which tries to overcome grammatical inconsistencies. Further there are two types of document in which the summarization can been applied- single document summarization where the important information from single document is selected and presented and the other is multi document summarization where the important information from a collection of documents is selected and presented.Also, considering the common language in India being Hindi, a summarizer for the same language is built.Therefore automatic summarization can be an important software for the people who do not know other languages other than Hindi. Hindi is written in the Devanagari script which has largest alphabet set. In the proposed system, the main goal is to summarize Hindi Text documents.

Earlier, text summarization was performed on single documents by the researcher. So here it states that text rank algorithm can be used for any language as it determines the important sentences from the paragraph irrespective of the meaning of the words. Also it is concluded that text rank gives better results for the summarization.

The main task is to remove the redundant sentences and the research on multi-document summarization focuses on the same. So it has been proposed to use text rank algorithm for summarizing documents in national language of India, Hindi.

2. LITERATURE SURVEY
Various authors have worked on the similar systems. Here is the detailed review of some of their work:

In a paper published by Vishal Gupta and Gurpreet Singh Lehal [1], they focus on feature selection and weight learning for Punjabi text summarization. The paper describes existing features such as sentence position, sentence length etc. and also describes new features such as common Punjabi-English noun feature, cue phrase feature and presence of e- mail address for sentence scoring and extraction.

In Extractive Text Summarisation in Hindi proposed by Sakshee Vijay, VartikaRai, Sorabh Gupta, AnshumanVijayvargia&DiptiMisra Sharma [2] uses extractive method and calculate the precision on big datasets which are gathered from 24253 news articles of hindi language and their extractive summaries evaluated with manual summaries of 60 words each.

In Automatic Summarization and Keyword Extraction from Web Page or Text File[3] the authors created an application which initially uses an algorithm which extract text from web pages and further apply methodologies to study page rank and text rank algorithm. Further, they have implemented knowledge to extract keywords using text rank algorithm.

In paper – Multi-Document Summarization Using Text Rank Algorithm and Maximal Marginal Relevance for Text in Bahasa Indonesia[4], the authors DaniGunawan, SitiHazizahHarahap, RomiFadillahRahmat have aimed to reduce the similar sentences which are obtained through

multi-document that share similar information and get an accurate text summary. Even though this process gives a summarized text, still there are some redundant similar sentences. To overcome this situation they have used Maximal Marginal Relevance (MMR) to reduce the similar sentences.

3. PROPOSED SYSTEM

The extraction technique of text summarization consists of selecting important sentences from source documents and arranges them in the destination documents. In this system areText Rank algorithm is used which is inspired from Page Rank.Page rank is primarily used for ranking webpages in online search results.

Now let's try to understand how Page Rank Algorithm works.

Table 1.Backlink matrix of website

Webpages	Links
P1	{P2, P3}
P2	{P1}
P3	{}

Here P1 has backlinks of {P2, P3}. P2 has backlinks of P1. P3 does not have any backlink. Now Page Rank Algorithm gives score to these webpages called Page Rank Score.

Now the probability of going from one page to other is given by – 1 / Unique links in the webpage. If there are no unique links then the probability is initialized to 0.

From above example, the Matrix obtained is :

Table 2. Resultant matrix obtained

	P1	P2	P3
P1	0	0.5	0.5
P2	1	0	0
P3	0	0	0

Hence these probabilities are used for ranking pages in online search results.

Now let's understand the approach of Text Rank algorithm. The following steps to be used in our approach –

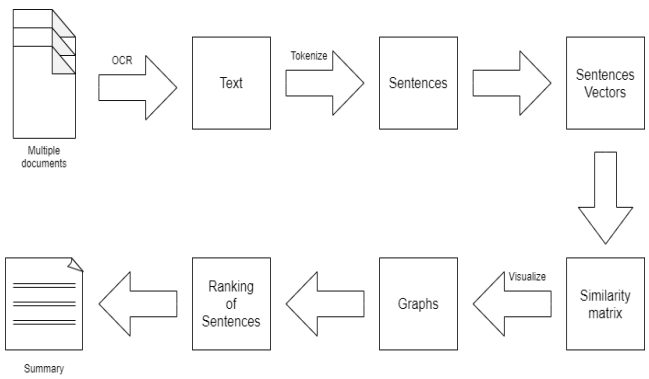

Fig 1: Flow diagram of the process

- From all the documents the sentences are extracted.
- The whole content is later split into sentences.
- The next task is to find vector representation of every sentence.
- Using similarity matrix, the similarity between two sentences is found out.
- Later the similarity matrix is visualized to graph, where sentences are taken up as vertices and similar scores are taken up as edges, from which sentence rank is calculated.
- From the obtained ranking of sentences the summary can be concluded.

The function of every step of the system is explained as below.

3.1 Pre-processing Step

In pre-processing phase,given set of images, first Thresholding is to be applied to the image so as to extract the text concisely. Then OCR(Optical Character Recognition) is applied. The text extracted is now been pre-processed. The Hindi text document is broken into sentences, sentences are further broken into words and then stop words are removed. Pre-processing involves preparing text document for the analysis and involves steps such as Tokenization, Stop words removal, Thresholding.

3.1.1 Tokenization: The text extracted from the document has paragraphs but for processing with the text, the text needs to be divided into small parts, that is, sentences. So tokenize it in sentences and then words as required.

3.1.2 Stop words removal: These are some commonly used words in hindi language. For text processing, these words need to be removed. अंदर, अत, अपना, etc are stop words which are to be filtered out beforehand.

3.2 Vector Representation of Sentences:

Word vectors are simply vectors of numbers that represent meaning of the word. Each sentence is given a score based on the weight of feature terms which is then used for sentence ranking. Feature term values ranges between 0 to1. There is one more possibility of using TF-IDF features for our sentences, but this method tends to ignore the order of placing the words. This might create meaningless sentences.

Hence use vectors from Gensim Model to form sentence vectors by using following formula:

Sentence Vector = (Sum of word vectors in that sentence) / (Length of sentence)

Now form a similarity matrix, for that a similarity between sentences. To do cosine similarity values need to be found out.

3.3Applying Text Rank Algorithm

The conversion of similarity matrix is done into the graph. In the graph,

similarity scores between the sentences. Then on the graph we will apply Text Rank Algorithm (i.e., obtained from Page Rank algorithm) .

After applying the algorithm sentence ranking is obtained and based on those ranking one can get top N sentences for summary formation.

4. RESULT AND ANALYSIS

Precision is number of correct sentences divided by no of sentences extracted. Recall is number of correct sentences divided by no of sentences that should have been extracted. F-score is mean of precision and recall. Well all the above modules are important and necessary to check accuracy, relevance and usefulness in the system which is implemented by Text Rank Algorithm. For testing purpose we even create a Manual summary and compared it with the summary obtained through our model to check for accuracy.

5. CONCLUSION

This system uses extractive method for multiple Hindi documents summarization. Further Text Rank Algorithm is used in this system to improve the quality of summarization. The summary generated by the system is found very close to summary generated by humans. The Precision, Recall & F-score values shows very good accuracy of summary generated by the system.

6. ACKNOWLEDGMENTS

Presentation, inspiration and motivation have always played a key role in the success of any venture.We would like to express our sincere gratitude towards our project mentor for continuous support in the study and research, for patience, motivation, enthusiasm and immense knowledge. Mentor guidance helped us in all the time of this project.We are pleased to present our project named “Hindi Text Summarization” and take this opportunity to express our profound gratitude to all those people who helped us in completion of this project.

7. REFERENCES

[1] Vishal Gupta, Gurpreet Singh Lehal, “Features Selection and Weight learning for Punjabi Text Summarization”, International

[2] Journal of Engineering Trends and Technology, Volume 2, Issue 2, 2011

[3] S. Vijay, V. Rai, S. Gupta, A. Vijayvargia and D. M. Sharma, "Extractive text summarisation in hindi," *2017 International Conference on Asian Language Processing (IALP)*, Singapore, 2017, pp. 318-321, doi: 10.1109/IALP.2017.8300607.

[4] X. You, "Automatic Summarization and Keyword Extraction from Web Page or Text File," 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET), Beijing, China, 2019, pp. 154-158, doi: 10.1109/CCET48361.2019.8989315.

[5] D. Gunawan, S. H. Harahap and R. Fadillah Rahmat, "Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia," 2019 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, 2019, pp. 1-5, doi: 10.1109/ICISS48059.2019.8969785.

[6] Patil Pallavi D, Mane P. M. , “A Comprehensive Review on Fuzzy Logic & Latent Semantic Analysis Techniques For Improving the Performance of text summarization”, International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS), Volume 2,Issue 11 Nov 2014.

[7] Pallavi D Patil, N. J. Kulkarni, “Text summarization using fuzzy Logic”, International Journal of Innovative Research in Advanced Engineering (IJIRAE),Volume 1, Issue 3, May 2014.

[8] S. Santhana Megala, A. Kavitha, A. Marimuthu , “Enriching Text Summarization using Fuzzy Logic”, International Journal of Computer Science and Information Technologies, Volume 5, Issue 1, 2014.

[9] S. A. Babar, S. A. Thorat, “Improving Text Summarization using Fuzzy Logic & latent Semantic Analysis”, International Journal of Innovative Research in Advanced Engineering, (IJIRAE),Volume 1, Issue 4, May 2014. K. Elissa, “Title of paper if known,” unpublished.

[10] Arman Kiani –B, M. R. Akbarzadeh, “Automatic Text Summarization Using Hybrid Fuzzy GA-GP”, IEEE International Conference on Fuzzy Systems, July 16-21, 2006.