HINDI TEXT SUMMARIZATION USING NUMERICAL METHODS

By

NISHANT KUMAR (CSE/14017)
ABHISHEK KUMAR (CSE/14008)

Bachelor Thesis submitted to

Indian Institute of Information Technology Kalyani

for the partial fulfillment of the degree of

Bachelor of Technology
in
Computer Science and Engineering

May, 2018

Indian Institute of Computer Technology

Department of Computer Science and Technology
Webel IT Park Campus, West Bengal – 741235

Certificate

This is to certify that the thesis entitled **"Hindi Text Summarization Using Numerical Methods"** being submitted by **Nishant Kumar, Abhishek Kumar** an undergraduate students (Reg. No 10114017, 10114008) in the Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal, India, for the award of Bachelors of Technology in Computer Science and Engineering is an original research work carried by them under my supervision and guidance. The thesis has fulfilled all the requirements as per the regulations of Indian Institute of Information Technology Kalyani and in my opinion, has reached the standards needed for submission. The work, techniques and the results presented have not been submitted to any other University or Institute for the award of any other degree or diploma.

Supervisor
(**Dr. Sanjay Chatterji**)
Assitant Professor
Department of CSE,
IIIT Kalyani

Co-Supervisor
(**Dr. Manjira Sinha**)
Visiting faculty from Industry
Department of CSE,
IIIT Kalyani

Date : Friday 4th May, 2018

Indian Institute of Computer Technology

Department of Computer Science and Technology
Webel IT Park Campus, West Bengal – 741235

Declaration

We hereby declare that the work being presented in this thesis entitled,**"Hindi Text Summarisation Using Numerical Method"**, submitted to Indian Institute of Information Technology Kalyani in partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering during the period from July, 2017 to April, 2018 under the supervision of **"Dr. Sanjay Chatterji"**, Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, India, does not contain any classified information.

Nishant Kumar
Reg No: CSE/10114017
Department Name : CSE
Institute Name :IIIT Kalyani

Abhishek Kumar
Reg No: CSE/10114008
Department Name : CSE
Institute Name :IIIT Kalyani

Acknowledgment

We take this opportunity to express our gratitude and regards to our guide **Dr. Sanjay Chatterji** for his exemplary guidance, monitoring and constant encouragement throughout the course of this project.
We also take this opportunity to express a deep sense of gratitude to our friends for their support and motivation which helped us in completing this task through its various stages.
We are obliged to the faculty members specially our co-guide **Dr. Manjira Sinha** visiting faculty from Industry of the Department of Computer Science and Engineering at "Indian Institute of Information Technology, Kalyani" for the valuable information provided by them in their respective fields. We are grateful for their cooperation during the period of my assignment.
Lastly, We thank our parents for their constant encouragement without which this assignment would not have been possible.

Nishant Kumar
Reg No: CSE/10114017
Department Name: CSE
Institute Name: IIIT Kalyani

Abhishek Kumar
Reg No: CSE/10114008
Department Name: CSE
Institute Name: IIIT Kalyani

Abstract

Automatic text summarization is a process which filters out the most essential part of the original source text/s. It eliminates the redundant, less important content and provides you with the vital information in a shorter version usually half a length of original text.
In our approach, we have used an extractive summarization method consists of selecting important sentences from the original document and concatenating them into shorter form.
In our proposed system, we are using numerical methods to extract summary which has an advantage that we don't need any previous data for summary extraction. The importance of sentences is decided based on score of numerical features of sentences like TF-ISF, Sentence Length, Sentence Position, Sentence Similarity, Numerical data.
Sentences will be selected depending on the scores gained by the sentences based on the several features. Higher the score of sentences, greater are the chances that they would be picked up in a summary. These scores are calculated on the basis of feature extraction for each sentence.
Hindi is taken as a study language for the proposed work.

Keywords: Hindi text summarization, extractive summarization, feature extraction, TF-ISF, Sentence Similarity

Contents

List of Figures

List of Acronyms

TF	Term Frequency
ISF	Inverse Sentence Frequency
ANN	Artificial Neural Network
SP	Sentence Position
∩	Intersection
HMM	Hidden Markov Model
SL	Sentence Length
ND	Numerical Data
DUC	Document Understanding Conferences

Chapter 1

Introduction

1.1 Objective

Internet exchanges a huge amount of data. Since last few years, Internet is being proliferated. So the problem of information overload has increased and hence the research in automatic summarization is increased too. Instead of reading the whole document that consists of many examples, comparisons, supported details, for readers it is always convenient to read point to point specific gist of the document. Hindi Automatic text summarization is exactly meant for the same. It provides the reader with filtered description of source text and a non-redundant presentation of facts found in the text.

1.2 Why Hindi is challenging ?

Although the various approaches are proposed for summarization of major languages like English,Swedish and other European languages. some challenging problems are still open for other languages of the world.
India has around 120 major languages and 1500 other languages.It the native language of most people living in Delhi, Chhattisgarh, Himachal Pradesh, Chandigarh, Bihar, Jharkhand, Madhya Pradesh, Haryana, and Rajasthan. While performing related search, it is observed that since recent years many researchers are working on Indian and other languages but less work has done on Indian languages.

For English, lots of libraries are already available for preprocessing and processing the document while for Hindi language there is not.
Though Hindi is the top-most language used in India and also in a few neighboring countries there is a lack of proper summarization system for Hindi text. Hence, the technique for Hindi text summarization has been proposed in this paper.

1.3 Application

Business leaders, analysts, students and academic researchers need to go through huge numbers of documents every day to keep ahead, and a large portion of their time is spent just figuring out what document is relevant and what isn't. By extracting important sentences and creating comprehensive summaries, it's possible to quickly assess whether or not a document is worth reading.

Usage

- Headlines of news.
- Abstract summary of technical paper.
- Review of book or Movie.

1.4 Organisation of Thesis

The rest of the paper is structured as follows:
Section 2 describes the related work in text summarization, and **Section 3** describes the proposed summarization technique. **Section 4** presents the experimental work and its results and **Section 5, 6, 7** concludes the paper by summarizing the study and gives some future work.

Chapter 2

Background and Literature Review

In this section, we present brief survey on variety of text summarization methods that has been proposed and evaluated, but mostly for English and European languages.

2.1 Related Work

Many previous works on extractive summarization uses mainly two major steps:(1) ranking the sentences based on the score which are computed by combining few or several features such as term frequency(TF), position information and cue phrases(Baxendale, 1958);(Luhn, 1958) and (2)selecting few top ranked sentences to form summary.
The very first work in automated text summarization was done by (Luhn, 1958). He used word frequency(number of times a word occours in a document) and phrase frequency as features to produce summaries. It has been assumed that the most frequent words are indicative of the main topic of a document.

Although subsequent research has devoloped several summarization methods based on the new features, the work presented by (Baxendale, 1958) is still used today as a foundation of extraction based summary. P.B. Baxendale (Baxendale, 1958) proposed novel feature that is sentence location or sentence location position in an input document. It is analyzed that sen-

tences which are positioned at the beginning or at the end of the document are more important than other sentences in the document.

H.P. Edmundson (Edmundson, 1969) proposed a novel structure for a text summarization. They proposed two new features. First, cuewords that is presence of most indicative words into a document such as finally, in summary, lastly, etc. Second, straightforward feature title or heading words for which an additional weight was assigned to a sentence, if sentences have heading words in it.

Later, (Kupiec, Pedersen, & Chen, 1995) proposed a machine learning approach to text summarization. They described a new technique of summarization with naive-Bayes classifier, the classification function classifies each sentence as whether it is extraction worthy or not.

(Conroy & O'leary, 2001) proposed two techniques for sentence extraction named QR(QR Matrix decomposition) and HMM(Hidden Markov Models). In QR method the importance of each sentence were measured and the most important sentence were added to the summary. Once this is done, the relative importance of the remaining sentences changed, because some of them were redundant. They repeated this process until they have captured enough of the important ideas.
The other technique was HMM which is a sequencial model for automatic text summarization that judges the likelihood that each sentence should be contained in the summary. In HMM only three features were used: sentence position, total no of terms in the sentence and similarity of sentence terms in the given document. In the end, evaluation was done by comparing HMM generated summary with human generated summary.

(Erkan & Radev, 2004) proposed a stochastic graph-based method for computing relative importance of textual units. They used LexRank approach for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences.

Chapter 3

Text Summarization and Extraction Techniques

3.1 Definition

Automatic text summarization is a technique which compresses large text into a shorter text which includes the important information.
The computer program is given a text and it returns a summary of the original text. This is done by reducing redundancy of the text and by extracting the essence of the text.

3.2 Types of Automatic Summarization

Automatic summarization can be performed in two different approaches(Wikipedia contributors, 2018):

1. Extraction-based Summarization
    I Statical Method
    II Linguistic Method
    III Hybrid Method
2. Abstraction-based Summarization

1. Extraction-based Summarization: This approach is to construct the summary by producing the most important sentences verbatim out of the original document and is mainly concerned with what the summary content should be.

2. Abstraction-based Summarization: The abstraction approach is to form summary by paraphrasing sections of the original document putting strong emphasis on the form, aiming to produce an important material in a new way.

3.3 Types of Extraction-based Methods

I **Statistical Method**: Text summarization based on this approach relies on the statistical distribution of certain features and it is done without understanding whole document. Models rank the sentences of the original text to appear in the summary in the order of importance. We are using average TF-ISF, title Word, sentence length, sentence feature, thematic word and numerical data as statistical features in our proposal.

II **Linguistic Method**: In this, method needs to be aware of and know deeply the linguistic knowledge, so that the computer will be able to analyse the sentences and then decide which sentence to be selected. We are using proper noun feature and sentence to sentence similarity as linguistic features in our proposal.

III **Hybrid Method**: It is the combination of statical and Linguistic methods.It optimizes best of both the previous methods for meaningful and short summary.

Chapter 4

Proposed model

The goal of automatic text summarization is to select the most important sentences of the Hindi text document. The proposed method uses various statistical features to find most important sentences. The general outline of the methodology used for this task is described below.

4.1 Flow of proposed system

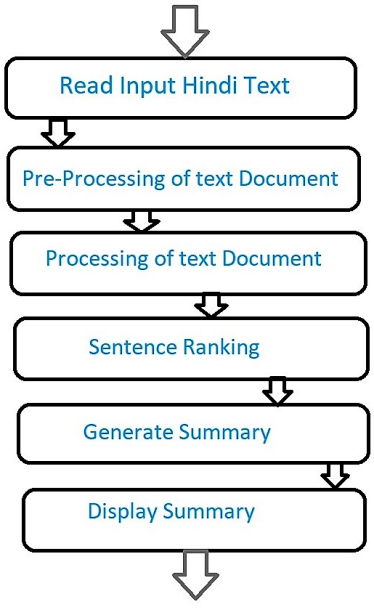


Figure 4.1: Flowchart of proposed model

Our proposed technique can be grouped into two major steps named as Preprocessing Step, Processing Step and Extraction Step which are explaind below in details.

4.2 Model Overview Steps

Input: A Hindi text file "Original-O" (Hindi.txt)
Output: A summarized text(S) of original text document(Hindi.txt) as per compression ratio.

1. Read Input text file.
2. Preprocess the file. (Processing step)
    I Sentence Segmentation.
    II Tokenisation of each sentence to words.
    III StopWords Removal.
    IV Stemming of Words.
3. Processing Step (feature extraction + sentence ranking)
    I Extract following features from "O" file
        i. Average TF-ISF
        ii. Sentence Length
        iii. Sentence Position
        iv. Sentence Similarity
        v. Numerical Data
    II Sentence Ranking to rank sentences in range of "0 to 1" with 1 indicating most important and 0 indicating not important sentence based on the each feature's normalised scores.
4. Generate Summary (Extraction Step)
    I While (Sentences in Summary file "S" does not exceed maximum limit as per given by compression ratio) extract all sentences from "S" sort by sentence rank of Maximum rank (rank5) to minimum (rank1).
5. Display Output Summary "S"

4.3 Pre-Processing Step

We need to prepare data for further processing. This intermediate preparation stage is called a Preprocessing step which is a structured representation of the original text.
In the preprocessing step of our proposed technique the input text obtained from the text file first split into sentences using segmentation method, then sentences are furthur split into words using tokenization and then stop-words are removed to clean the original text.
Our proposed pre-processing step consist of four steps 1) Sentence Segmentation 2) Tokenization 3) Stop-Words Removal and 4) Stemming.

4.3.1 Sentence Segmentation

It is boundary detection for a sentence. The purpose of segmentation is to use sentence segments as a basic unit that possibly conveys independent meanings.
In Hindi, sentence is segmented by identifying boundary of sentence that ends with purnaviram(|).

4.3.2 Tokenization

In tokenization the sentences are broken up into discrete bits or tokens(words). It omits certain characters, such as punctuation, spaces and special symbols between words. Punctuations(Viram Chinha) in Hindi language consists of UpViram(:), ArdhViram(;), PurnaViram(|)etc.

4.3.3 Stop-Words Removal

Stop-Words include function words, articles, prepositions, conjunctions, prefix, postfix, etc. i.e. common words that carry less important meaning than keywords.
So these types of words should be removed from input text document, otherwise the sentence having more no of stop-words could have higher weight. We analyzed that every Hindi text document contains minimum

25-30%. Also they make the text look heavier and are insignificant. Hence should be eliminated.
For example:

STOP WORD EXAMPLES

कारक	ने, को, से, के लिए, में , पर	विशेषण	थोड़ी, कुछ, कौन,अनेक
समुच्चय बोधक अव्यय	और,लेकिन , पर,एंव, इसलिए,मगर	समास	अनुसार , पर्यन्त, वाला
विस्मयादि बोधक अव्यय	अहा! शाबाश! हाय! अरे! हट!	अव्यय	धीरे– धीरे, बहुत,तथा, तक,ही , भी
सर्वनाम	आप, तू ,यह , वह ,कूछ		

Figure 4.2: Stop-Words Example in Hindi

4.3.4 Stemming

In Stemming process, the suffixes are ignored and removed from words to get the common origin. It recognizes words with common meaning and form as being identical. Syntactically similar words, such as plurals, verbal variations, etc. are considered similar. e.g. “walk”, “walking” and “walked” are counted as same and derived from a stem word “walk”.

4.4 Processing Step

In processing step, we decide and calculate the features that affect the relevance of sentences and then weights are assigned to these features using weight learning method. Higher ranked sentences are extracted for summary.
Feature Extraction: Real analysis of the document for summarization begins in this phase. Every sentence is represented by the feature terms vector and has a score based on the weight of feature terms. This score is used for sentence ranking. Feature term values range between 0 to 1. Six statical features are used as follows:

4.4.1 Average TF-ISF

TF-ISF stands for term frequency-inverse Sentence frequency and the TF-ISF weight is a numerical measure used to evaluate how important a word is to a document. The importance increases proportionally to the number of times a word appears in the sentence (TF) but is offset by the frequency of the word in the corpus (ISF).

$$TF(t,S) = \frac{\text{No of times term } \mathbf{(t)} \text{ appears in a Sentence } \mathbf{S}}{\text{Total number of Words in the Sentence } \mathbf{d}}$$

We should look at the distribution of the word across the complete document instead of making only a local comparison. The intention is to punish a word that occurs frequently all over the text, but are little informative. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$ISF(t,S) = \log\left(\frac{\text{Total No. of Sentences}(|S|)}{\text{Number of Sentences containing the term } \mathbf{t}}\right)$$

The score of a sentence k is computed based on the frequncy of important words occourance in a sentence.

$$AvgTFISF(S_t) = \sum TF * ISF$$

4.4.2 Sentence Length(SL)

The short sentences such as datelines and author names are not expected to belong to the summary. In the same way, too long sentences may contain a lot of redundant data and hence are unlikely to be included in the summary. So, we eliminate the sentences which are too short or too long. This feature computation uses minimum and maximum length threshold values.

L = *Length of the sentence*
MinL = *Minimum length of sentence = (5 in our experiment)*
MaxL = *Maximum length of sentence = (15 in our experiment)*
Minθ= *Minimum Angle(0* deg*)*
Maxθ= *Maximum Angle(180* deg*)*
SL = 0, If (L $<$ MinL) or (L $>$MaxL)
Otherwise
SL = sinθ,(Student & COE, 2015)
Where,

$$[\theta = \frac{(\max\theta - \min\theta)}{(\max L - \min L)} \times (L - \min L)]$$

4.4.3 Numeric Data(ND)

Usually the numerical data is used to show the important mathematical or statistical analysis providing some vital information in a document and hence claims to be a part of summary with its essential contribution to the document. Thus the ratio of the number of numerical data in a sentence to the sentence length is used as a score for this feature.

$$ND = \frac{Number\ of\ Numeric\ Data\ in\ Sentence}{Sentence\ Length}$$

4.4.4 Sentence Position (SP)

Usually, sentences in the beginning defines the theme of the document, while end sentences conclude or summarize the document. So, position of the sentence in the text, decides its importance.
Threshold value in percentage, defines how many sentences in the beginning and at the end are retained in summary with weight SP=1

$$[SP = \cos(\frac{(\max\phi - \min\phi)}{(\max V - \min V)} \times (CP - \min V))], (Student\ \&COE, 2015)$$

Where, $THLD =$ Threshold Value(10% in our Experiment)

$MinV = NS \times THLD$(Minimum value of Sentence)

$MaxV = NS \times 1 - THLD$(Maximum value of Sentence)

$NS =$ Number of Sentences in document

$Min\phi =$ Minimum Angle(0 deg)

$Max\phi =$ Maximum Angle(360 deg)

$CP =$ Current Position of Sentence

4.4.5 Sentence to Sentence Similarity

For each sentence **S** compute the similarity between S by creating a weighted matrix with other sentences **S'** of the document, then add up those similarity values. It gives us the raw value of this feature for **S**. There are many approaches to calculate the similarity between two sentences.

$$SS = \sum_{1}^{N} Sim(i,j) When, i \neq j$$

$Sim(i,j) =$ Number of Words overlapping between sentence S(i) and S(j)

4.5 Sentence Ranking

Sentence ranking is the very important step to determine which sentence should be included in our summary.
In our proposed method sentence ranking per sentence is evaluated using individual score obtained from each feature per sentence and then overall collective score of each sentence is computed by adding up individual score of each features. Then each score of sentences is normalised between 0 to 1. Sentences are sorted based on the descending order of score values. Depending on the compression rate, sentences are extracted from the document to generate summary.

4.6 Generate Summary

Using ranking for each sentence obtained from above sentence ranking step, depending upon the user compression rate input all sentences are extracted till sentence in the summary file does not exceed maximum limit.Then summary is generated in the order of original text document.

4.7 Display Summary

Using generated summary based on the user compression ratio, summary is displayed on the screen as well as written into a text file.

Chapter 5

Summary Example

We have used a news dataset obtained from an article from “AmarUjala” News Paper in summary example.

Original Paragraph

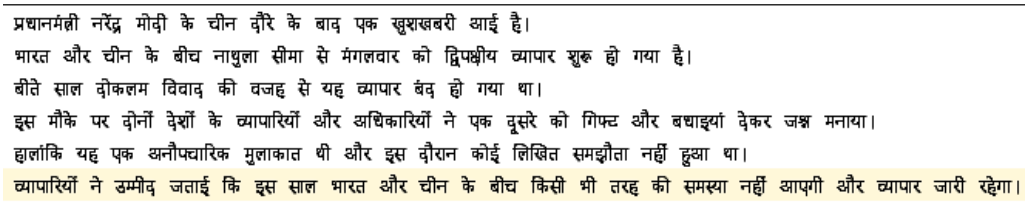
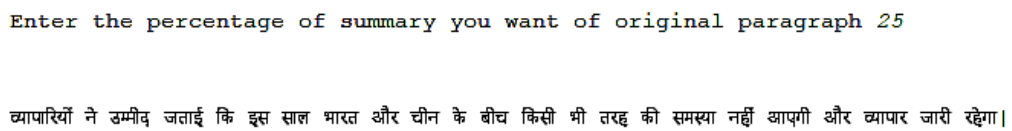

प्रधानमंत्री नरेंद्र मोदी के चीन दौरे के बाद एक खुशखबरी आई है।
भारत और चीन के बीच नाथुला सीमा से मंगलवार को द्विपक्षीय व्यापार शुरू हो गया है।
बीते साल डोकलम विवाद की वजह से यह व्यापार बंद हो गया था।
इस मौके पर दोनों देशों के व्यापारियों और अधिकारियों ने एक दूसरे को गिफ्ट और बधाइयां देकर जश्न मनाया।
हालांकि यह एक अनौपचारिक मुलाकात थी और इस दौरान कोई लिखित समझौता नहीं हुआ था।
व्यापारियों ने उम्मीद जताई कि इस साल भारत और चीन के बीच किसी भी तरह की समस्या नहीं आएगी और व्यापार जारी रहेगा।

Figure 5.1: Original Paragraph.

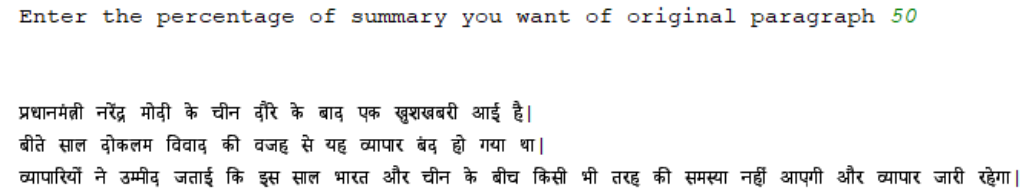
Paragraph after 25 percent compression ratio.


Enter the percentage of summary you want of original paragraph 25

व्यापारियों ने उम्मीद जताई कि इस साल भारत और चीन के बीच किसी भी तरह की समस्या नहीं आएगी और व्यापार जारी रहेगा।

Figure 5.2: Paragraph after 25 percent compression ratio.

Paragraph after 50 percent compression ratio


Enter the percentage of summary you want of original paragraph 50

प्रधानमंत्री नरेंद्र मोदी के चीन दौरे के बाद एक खुशखबरी आई है।
बीते साल डोकलम विवाद की वजह से यह व्यापार बंद हो गया था।
व्यापारियों ने उम्मीद जताई कि इस साल भारत और चीन के बीच किसी भी तरह की समस्या नहीं आएगी और व्यापार जारी रहेगा।

Figure 5.3: Paragraph after 50 percent compression ratio.

Chapter 6

Evaluation

To test our summarization system, we collected 20 Hindi documents from the Hindi daily newspaper, AmarUjala. The documents are typed and saved in the text files using UTF-8 format. For each document in our corpus, we consider only one reference summary for evaluation. Evaluation of a system generated summary is done by comparing it to the reference summary.

It is very difficult to determine whether a summary is good or bad. The summary evaluation methods can be broadly categorized as human evaluation methods and automatic (machine-based) evaluation methods. A human evaluation is done by comparing system-generated summaries with reference/model summaries by human judges.
The automatic evaluations may lack the linguistic skills and emotional perspective that a human has. Hence although automatic evaluation is not perfect compared to the human evaluation, it is popular primarily because the evaluation process is quick even if summaries to be evaluated are large in number. Since automatic evaluation is performed by a machine, it follows a fixed logic and always produces the same result on a given summary.

In several past Document Understanding Conferences (DUC) organized by NIST (The National Institute of Standards and Technology), single document text summarization systems for English have been evaluated. In DUC 2001 and DUC 2002, single document summarization task was to

generate a summary of fixed length such as 50 words, 100 words etc. A baseline called LEAD baseline was defined in theses conferences. LEAD baseline considers the first n words of an input article as a summary, where n is a predefined summary length.
Unlike DUC single document text summarization task where there was a fixed summary length for each document, we believe that a generic summary of a document may be longer or shorter than a summary of another document. So, we assume that the size of a system generated summary should be equal to that of the corresponding model summary, but the different model summaries may not be equal in size.
We adopted an automatic summary evaluation metric for comparing system-generated summaries to reference summaries. When we compare a system generated summary to a reference summary, we ensure that they would be of the same length. We have used the unigram overlap method stated in (Radev et.al, 2004) for evaluating the system generated summaries. Unigram overlap between a system generated summary and a reference summary is computed as follows:

$$\text{Unigram based Recall Score} = \frac{|R| \textbf{ intersection } |S|}{|R|}$$

$$\text{Unigram based Precision Score} = \frac{|R| \textbf{ intersection } |S|}{|S|}$$

$$\text{F measure Score} = 2 \times \frac{|R| * |S|}{|R| + |S|}$$

$|R|$ is the length of the reference summary,
and $|S \cap R|$ indicates the maximum number of unigrams co-occurring in the system generated summary S and the reference summary R.
Creation of reference summaries is a laborious task. In our experiment, we have used only one reference summary for evaluating each system generated summary.

Chapter 7

Conclusion and Future Work

This paper discusses a single document text summarization method for Hindi. Many techniques have been developed for summarizing English text(s). But, a very few attempts have been made for Hindi text summarization.

The performance of the proposed system may further be improved by improving stemming process, exploring more number of statical and linguistic features like proper noun and applying learning algorithm for effective feature combination. Traditionally, more than one reference summaries are used for evaluating each system generated summary, but in our work, we have used only one reference summary for summary evaluation. In future, we will consider more than one reference summaries for summary evaluation. In future, more features like named entity recognition, cue words, context information, world knowledge etc, can be added to improvise the technique. Also, same technique can be applied on domains other than news and later we can study the effects of various domain characteristics on the suggested features and overall performance of the technique.

It can also be extended to work on multiple documents. Also ANN(Artificial Neural Network) based approach can be used for sentence ranking.This method involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not.

References

Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. *IBM Journal of Research and Development*, *2*(4), 354–361.

Conroy, J. M., & O'leary, D. P. (2001). Text summarization via hidden markov models. In *Proceedings of the 24th annual international acm sigir conference on research and development in information retrieval* (pp. 406–407).

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264–285.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international acm sigir conference on research and development in information retrieval* (pp. 68–73).

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159–165.

Student, P., & COE, D. M. (2015). A comparative study of hindi text summarization techniques: Genetic algorithm and neural network.

Wikipedia contributors. (2018). *Automatic summarization — Wikipedia, the free encyclopedia.* Retrieved from `https://en.wikipedia.org/w/index.php?title=Automatic_summarization&oldid=822496672` ([Online; accessed 29-April-2018])