Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 4, Issue 1; January-March, 2017, pp. 1-4

http://www.krishisanskriti.org/Publication.html

Multi Document Hindi Text Summarization using Back Propagation Network

Ashlesha Giradkar[1] and S.D. Sawarkar[2]

[1]PG Student, Archana Gulati
[2]Archana Gulati
E-mail: [1]ashlesha554@gmail.com, [2]archana.gulati@gmail.com

Abstract—*Now days it's very difficult and laborious task to find out exactly want we want from internet. To make this task there are many summarization technique has been developed. For English language there are multiple options available but very less work has been done with respect to Hindi language. Proposed system is going to summarize multiple Hindi documents. In this summarization technique feature extracted from document such as sentence length, sentence position etc are used to calculate which sentence should be included in summary.*

1. INTRODUCTION

These days text summarization become more popular among researcher due to the problem of overloaded information on the web and so there is necessity of more powerful text summarization technique. Day by day information available online increases tremendously so it becomes tedious task for user to find out relevant information. There are two summarization technique text abstraction and text extraction. Extraction is where summary consists of sentences extracted from document and in abstraction the focus is on idea of document in short is in summary. Overall process of Text summarization is basically divided into three parts: preprocessing of text document, sentence scoring based on extraction and last is development of summary. Summarization technique helps in providing quick summary of information contained in the document.

2. CATEGORIZATION OF SUMMARIZATION

2.1 Abstract vs. Extract Summary

In abstraction process paraphrasing of source document takes place. Extraction is the process of picking subset of sentences from source document and presents them to user in the form of summary.

2.2 Generic vs. Query-Based summary

Generic summary addresses broad community of reader. Query based summary are formed according to specific needs of an individual or a particular group and represent particular topic.

2.3 Single vs. Multi-Document Summary

In single document summary provide most relevant information contained in single document whereas multi-document summary helps to identify redundancy across documents and compute the summary of a set of related documents of a corpus such that they cover the major details of the events in the document.

2.4 Indicative vs. Informative

An indicative summary provides indication of subject of document and in informative summary reflects the content and allows describing what was in the input text.

2.5 Background vs. Just-the-News

A background summary assumes the reader's prior knowledge of the general setting of the input text content is poor, and hence includes explanatory material, such as circumstances of place, time, and actors. Just-the-news summary contains just the new or principal themes, assuming that the reader knows enough background to interpret them in context.

3. PROPOSED SYSTEM

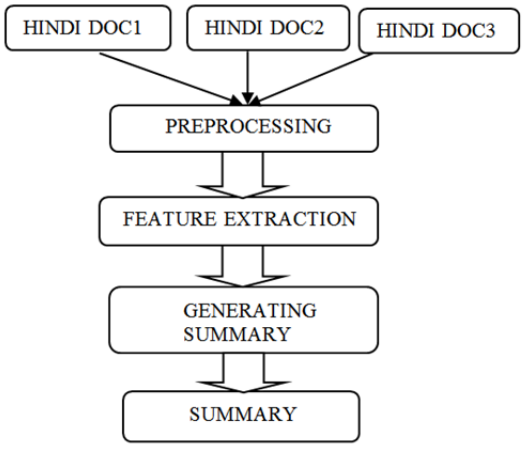

Fig. 3.1: Proposed system

The proposed methods find out most relevant sentence from multiple documents by using statistical approach. This summarization process has three major steps preprocessing, extraction of feature and implementation of backpropagation network.

3.1 Preprocessing

Preprocessing is nothing but the prepare source Document for analysis. This preparation is basically going to perform in four steps sentence segmentation, sentence tokenization, stop word removal and stemming.

3.1.1 Segmentation

In sentence segmentation step given text document is divided into sentence by sentence along with its word count. In Hindi language sentences are identify by purna viram(|).

3.1.2 Tokenization

In tokenization step sentence are divided into words by identifying spaces, comma and special symbols between words. So till now there is ready list of sentences with its words count for further processing.

3.1.3 Stop word removal

In stop word removal step some common words which do not aggregate relevant information to task are removed so that feature implementation use effectively by only considering words in the document which have more important. Stop words are common words that carry less important meaning than keyword, are eliminated for better summary generation.

3.1.4 Stemming

Stemming is process of obtaining root of each word which emphasizes its semantics. By these procedure syntactically similar words such as plurals, verbal variations etc. are considered similar. Stemming is used for matching words of sentences for checking similarity feature. Steamer used is developed by IIT Mumbai.

3.2 Feature extraction

In feature extraction step every sentence is represented by a vector of feature terms. Each sentence has a score based on the weight of feature terms which in turn is used for sentence ranking. Feature term values ranges between 0 to1. Following section describes the features used in this study.

3.2.1 Average TF-ISF (Term Frequency Inverse Sentence Frequency)

Term frequency is nothing but evaluation of distribution of word over document. Words which are common in sentences are less important when it's come to differentiate sentences. In other word it is important to calculate in how many sentences certain word exists which is nothing but inverse sentence frequency.

$$TF = \frac{\text{Word occurrence in sentence}}{\text{Total number of words in sentence}}$$

SF = Sentence frequency is count of sentence in which word occurred in a document of N sentence

ISF = log [Total Sentences / SF]

tf*isf = TF * ISF

Average tf*isf is calculated for each sentence and assigned as weight to the sentence.

3.2.2 Sentence length

Sentence length feature is used to filter out short or long sentence. If sentences are too long or too short then this kind of sentence are not good for summary. So in this feature maximum and minimum range for sentence is given.

SL = 0 if L < MinL or L > MaxL

Otherwise

$$SL = \frac{\text{Sin}((L-\text{MinL}) * ((\text{Max}\,\phi - \text{Min}\,\phi)}{(\text{MaxL} - \text{MinL}))}$$

Where, L = Length of Sentence

MinL = Minimum Length of Sentence

MaxL = Maximum Length of Sentence

Min ϕ = Minimum Angle (Minimum Angle=0)

Max ϕ= Maximum Angle (Maximum Angle=180)

3.2.3 Sentence position

In document position of sentence decide sentences importance. In most of document sentences in beginning indicates theme of document whereas sentences at end are indicates summary of document.

In Sentence position feature define threshold value in percentage which shows how many sentences in beginning and at end are retained in summary.

3.2.3 Numerical data

In numerical data feature sentences that contains

Number or digits are given more importance than

Other sentences. These sentences must be included

in summary.

ND=1, Digit exist

ND=0, Digit does not exist

3.2.4 Sentence to sentence similarity

In sentence to sentence similarity feature finds similarity between sentences. For each sentences similarity between that sentence and other sentences is computed by the method of stemmed word matching.

Sim(k,l)= Number of words occurred in sentence/wt

Where, wt=total words in sentence

3.2.5 Title feature

In document words of title represent main idea. So if sentence have higher intersection with title words then we can conclude that sentence is more important and must be part of summary generated.

3.2.6 SOV qualification

Sentences are formed by group of words expressing some thought so it must have a subject and verb. In Hindi language order of words are somewhat flexible but typical word order in most of the sentences is <subject><object><verb>. For checking SOV qualification of sentence, each word in sentence is tagged by assigning part of speech. Then based on tagged assigned, the first noun word in the sentence is marked as subject of sentence and then entire sentence is parsed till end, if verb is last word of sentence then sentence SOV qualified.

3.2.7 Subject similarity

In sentence similarity feature subject of sentence is matched with subject of title. For finding sentence subject result of previous step is used.

3.3 Generating Summary

In this step final generation of summary takes place. All summary generation completed in two phase network training and sentence selection. Here network used consist of simple three layer structure input layer, hidden layer and output layer. Input layer consist of 8 neurons this neuron takes input form 8 features as describe above. Hidden layer consist of 300 neurons. Output layer consist of single neuron.

3.3.1 Network training

The first phase of the process involves training

the backpropagation networks learn the types of sentences that should included in summary. this is done by traning network with sentences in several text document. Here every sentence has to be identified whether it included in summary or not and this is perform by human reader. Once training is completed network is ready.

backpropagation of error

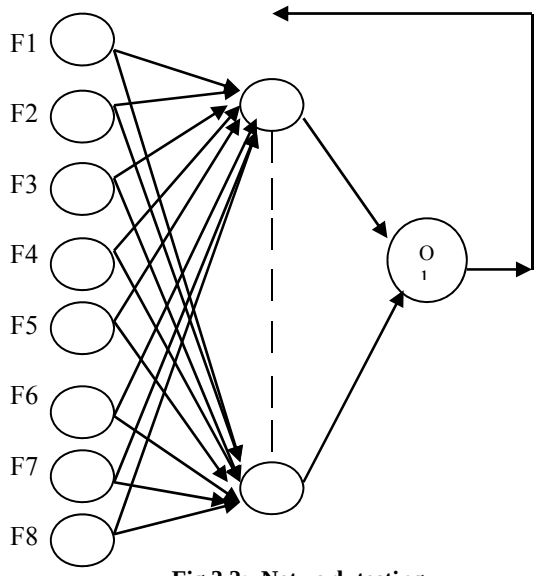

Fig 3.2: Network testing

3.3.2 Sentence selection

Once the network has been trained, it can be used to filter sentences in paragraph and determine whether each sentence should be included in the summary or not. This phase is accomplished by providing control parameters for hidden layer activation to select highly ranked sentences.

4. CONCLUSION

In this paper discussion is on Hindi text summarization using extractive method. An Extractive summary is selection of important sentences from Hindi text Documents. The importance of sentences is decided based on statistical and Linguistic feature of sentences. This summarization procedure is based on backpropagation network.

REFERENCES

[1] Chetana Thaokar and Dr Latesh Malik, "test model for summarizing hindi text using extraction Method", ieee conference on ict 2013.

[2] K.Vimal Kumar and Divakar Yadav, "an improvised extractive approach to hindi text summarization", Springer, india 2015.

[3] Vishal Gupta and Gurpreet Singh Lehal, " a survey of text summarization extractive techniques", Journal of emerging technology in web intelligence, vol 2, no 3, aug 2010

[4] Khosrow Kaikhah" automatic text summarization with neural networks" second ieee international conference on intelligent systems, june 2004

[5] Mr. Sarda a.t., Mrs. Kulkarni a.r. "text summarization using neural networks and rhetorical structure theory" international journal of advanced research in computer and communication engineering vol. 4, issue 6, june 2015