International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), September 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

A MULTI-DOCUMENT HINDI TEXT SUMMARIZATION TECHNIQUE USING FUZZY LOGIC

Arti S.Bhoir[1], Archana Gulati[2]

[1,2]University of Mumbai, (INDIA)

ABSTRACT

Today it is very difficult, laborious and time consuming task to extract information manually from large amount of data available. Many summarization techniques have been developed for English language but very less research work is carried out in the area of Hindi text summarization. Until now we have done Hindi text summarization on single source document. In the proposed system, the idea is to summarize multiple Hindi text documents, using sentence extractive method. This summarization is based on features extracted from documents such as sentence length, sentence position etc. Some new features namely Hindi cue phrase, Hindi-English common words, presence of URL's or email addresses in sentences have been introduced. Thus, the proposed system can be used for Hindi text summarization of multiple documents based on Fuzzy logic system.

Keywords: *Hindi text summarization, text summarization techniques, sentence features fuzzy logic.*

I. INTRODUCTION

A Text summarization is one of the most popular research areas today because of the problem of the information overloading available on the web, and has increased the necessity of the more strong and powerful text summarizers. Natural language may be English, Hindi, Marathi, or any other regional languages as opposed to artificial languages, like C++, Java. Text summarization is one of the applications of Natural Language Processing.

With the increasing amount of online information, it becomes extremely difficult to find relevant information to users. Information The single document summarizer is an application which is proposed to extract the most important information of the document. In automatic summarization, there are two distinct techniques either text extraction or text abstraction. Extraction is a summary consisting of a number of sentences selected from the input document. An abstraction based summary is generated where some text units are not present into the input document. The total system is alienated into three segments: pre-processing the text document, sentence scoring based on text extraction and summarization based on sentence ranking [2]. Retrieval systems usually return a large amount of documents listed in the order of estimated relevance. It is not possible for users to read each document in order to find useful ones. Automatic text summarization systems helps in this task by providing a quick summary of the information contained in the document. An ideal summary in these situations will be one

that does not contain repeated information and includes unique information from multiple documents on that topic.

In the proposed system, the idea is to summarize multiple Hindi text documents, using sentence extractive method [1]. This summarization is based on features extracted from documents such as sentence length, sentence position, Some new features namely Hindi cue phrase, Hindi-English common words, presence of URL's or email addresses in sentences have been introduced. Thus, the proposed system can be used for Hindi text summarization of multiple documents based on Fuzzy logic system [13].

II. SUMMARIZATION APPROACHES

There are different types of summarization approaches depending on what the summarization method focuses on to make the summary of the text.

2.1 Statistical Approaches

In Statistical methods, sentence selection is done based on word frequency, indicator phrases and other features regardless of the meaning of the words. These methods are based on the idea that text surface cues are the most obvious indication of the text contents. There are several methods for determining the key sentences such as, the title method, the location method, the aggregation similarity method, the frequency method etc. Automatic text summarization system in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights [4]:

2.1.1 Cue Method

This is based on the hypothesis that the relevance of a sentence is computed by the presence or absence of certain cue words in the cue dictionary.

2.1.2 Title Method

Here, the sentence weight is computed as a sum of all the content words appearing in the title and (sub-) headings of a text.

2.1.3 Location Method

This method is based on the assumption that sentences occurring in initial position of both text and individual paragraphs have a higher probability of being relevant. The results showed, that the best correlation between the automatic and human-made extracts was achieved using a combination of these three latter methods [1].

2.2 Linguistic Approaches

In this, method needs to be aware of and know deeply the linguistic knowledge, so that the computer will be able to analyses the sentences and then decide which sentence to be selected. It identifies term relationship in the document through part-of-speech tagging, grammar analysis, thesaurus usage and extracts meaningful sentences. Parameters can be cue words, Title feature or Noun and verbs in the sentences Linguistic approaches are based on considering the connections between words and trying to find the main concept by analyzing the words. There are some methods such as, Lexical chain, Graph theory, Clustering etc [1].

2.3 Hybrid Method

It exploits best of both the previous method for meaningful and short summary.

III .CATEGORIZATION OF SUMMARIZATION

3.1 Abstract vs. Extract Summary

Abstraction is the process of paraphrasing sections of the source document whereas extraction is the process of picking subset of sentences from the source document and presents them to user in form of summary that provides an overall sense of the documents content [5].

3.2 Generic vs. Query-Based summary

Generic summary do not target to any particular group. It addresses broad community of readers while Query or topic focused queries are tailored to the specific needs of an individual or a particular group and represent particular topic [5].

3.3 Single vs. Multi-Document Summary

single document summary provide the most relevant information contained in single document to the user that helps the user in deciding whether the document is related to the topic of interest or not 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 documents, taking into account some of the major issues : the need to carefully eliminate redundant information from multiple documents and achieve high compression ratios; information about document and passage similarities, and weighting different passages accordingly; the importance of temporal information; co-reference among entities and facts occurring across documents [5].

3.4 Indicative vs. Informative

An indicative summary provides merely an indication of the principal subject matter or domain of the input text(s) without including its contents. After reading an informative summary, one can explain what the input text was about, but not necessarily what was contained in it. An informative summary reflects (some of) the content, and allows one to describe (parts of) what was in the input text.

3.5 Background vs. Just-the-News

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

IV. PROPOSED SYSTEM

The proposed method uses statistical and linguistic approach to find most relevant sentences from multiple documents [1]. As shown in Fig.4.1, summarization system consists of three major steps preprocessing, extraction of feature terms & fuzzy logic for ranking the sentence based on the optimized feature weights.

4.1 Preprocessing Step

Preprocessing, involves preparing text document for the analysis. This step involves Sentence segmentation, Sentence tokenization, Stop word Removal and Stemming [1].

4.1.1 Sentence Segmentation

It is the process of decomposing the given text document into its constituent sentences along with its word count. In Hindi, sentence is segmented by identifying the boundary of sentence which ends with purna viram (|) .

4.1.2 Tokenization

It is the process of splitting the sentences into words by identifying the spaces, comma and special symbols between the words. So list of sentences and words are maintained for further processing.

4.1.3 Stop Word Removal

To effectively use word feature score we need to only consider the words in the document which have importance**. C**ommon words with no semantics and which do not aggregate relevant information to the task are eliminated / removed. Stop words are common words that carry less important meaning than keywords .This words should be eliminated otherwise sentence containing them can influence summary generated.

4.1.4 Stemming

Syntactically similar words, such as plurals, verbal variations, etc. are considered similar, the purpose of this procedure is to obtain the stem or root of each word, which emphasize its semantics. For e.g. Foxes, a root word of foxes is fox. Stemming is used for matching the words of sentence for checking similarity features. Stemmer used is developed by IIT Mumbai.

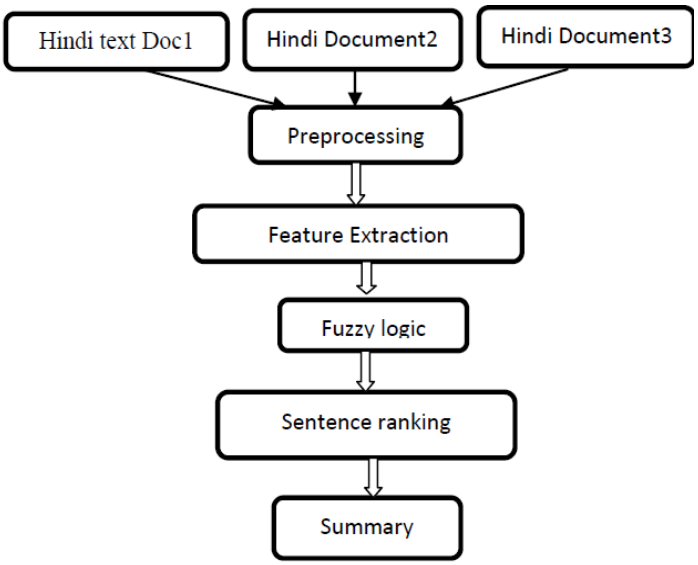

Fig.4.1. Proposed System

4.2 Feature Extraction

In this every sentence is represented by an vector of feature terms. This checks for every sentence statistically and linguistically. Each sentence has a score based on the weight of feature terms which in turn is used for sentence ranking [1]. Feature term values ranges between 0 to1.

Following section describes the features used in this study.

F1: Average TF-ISF (Term Frequency-Inverse Sentence Frequency):

TF means to evaluate distribution of each word over the document. Inverse sentence frequency means the terms that occur in only a few sentences which are more important than others that occur in many sentences

of the document. In other words, it is important to know in how many sentences a certain word exists. Since a word which is common in a sentence, but also it is common in the most of the sentences that is less useful when it comes to differentiating that sentence from other sentences.

This feature is calculated as "(1)"

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

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

ISF = log [Total Sentences / SF]

tf*isf = TF * ISF

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

F2: Sentence Length

This feature is useful to filter out short or long sentences. Too short or long sentence is not good for summary. This feature computation uses minimum and maximum length threshold values. The feature weight is computed as "(2)".

SL = 0 if L < MinL or L > MaxL (2)

Otherwise

SL = Sin ((L-MinL) * ((Max Θ- Min Θ)/ (MaxL - MinL)))

Where, L = Length of Sentence

MinL = Minimum Length of Sentence

MaxL = Maximum Length of Sentence

Min Θ = Minimum Angle (Minimum Angle=0)

Max Θ = Maximum Angle (Maximum Angle=180)

F3: Sentence Position

Position of the sentence in the text, decides its importance. Sentences in the beginning defines the theme of the document whereas end sentences conclude or summarize the document. In this threshold value in percentage defines how many sentences in the beginning and at the end are retained in summary whose weight is given as "3",SP = 1.

Remaining sentences, weight is computed as follows

Sp = Cos ((CP - MinV)*((MaxΘ - MinΘ) / (MaxV - MinV)))

Where TRSH = Threshold Value

MinV = NS * TRSH (Minimum Value of Sentence)

MaxV = NS * (1 - TRSH) (Maximum Value of Sentence)

NS = Number of sentences in document

MinΘ = Minimum Angle (Minimum Angle=0)

MaxΘ = Maximum Angle (Maximum Angle=180)

CP = Current Position of sentence

F4: Numerical Data

The Sentence that contains numerical data is important and it should be included in the summary. The Weight for this feature is calculated as "4".

ND = 1, Digit exist

0, Digit does not exist

F5: Sentence to Sentence Similarity

This feature finds the similarity between the sentences. For each sentence S, similarity between S and every other sentence is computed by the method of stemmed word matching.

Sim (i, j) = Number of words occurred in Sentences (Sj)

WT

Where, N = Number of Sentences

WT = Total Words in Sentence Si

Individual sentence weight based on similarity is the ratio of SS to N-1.

F6: Title Feature

Title contains set of words that represents gist of the document. So if a sentence Si has higher intersection with the title words then we can conclude Si is more important than other sentences in that document. Title score is calculated as "6"

TS (Si) = Number of words occurred in title

WT

F7: SOV Qualification

Sentence is a group of words expressing a complete thought, and it must have a *subject* and a *verb*. The word order in Hindi is somewhat flexible. However, the typical word order of the most of the sentences is <subject> <object> <verb>. For this reason, Hindi is sometimes called an "SOV" language. For SOV qualification of a sentence, each word in a sentence is tagged by assigning part of speech like (Noun, Adjective ,Verb, Adverb). Now based on the tags assigned, the first noun word in the sentence is marked as subject of the sentence. Whole sentence is parsed till its end, if verb is last word of the sentence than sentence is qualified as SOV. Only those sentence which are qualified as SOV will be used for further processing. Sentence after removing stop word is used.

F8: Subject Similarity

For subject similarity feature, a result of previous step is used to match subject of the sentence with the subject of the title. It can be similar to noun checking of title and sentence. Noun plays an important term in understanding the sentence. It is given as "8".

Sub (Si) = 1, if POS is noun and root value

Of title and sentence is equal

0, otherwise

F9: Hindi Cue Phrase Feature

Cue Phrases are certain keywords like In conclusion, summary and finally etc. These are very much helpful in deciding sentence importance. Those sentences which are beginning with cue phrases or which contain these cue phrases are generally more important than others [10]. Firstly a list of Hindi Cue phrases has been made and then those sentences containing these cue phrases are given more importance

F10: Hindi-English Common Words

Hindi sentences may contain some Hindi-English common words such as technology etc. Sentences containing Hindi-English common words are important and have higher probability to be extracted for the summary [10]. Hindi-English common words are calculated using equation 7.

HindiEnglish (si) = HinEng (si) SenLen (si)

Where HinEng (*Si*) = Total no of Hindi-English common words in sentence *Si*.

F11: Presence of URL's or Email Addresses

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 [10].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.

4.3 Fuzzy Logic

In this System, we propose important sentence extraction using fuzzy rules and fuzzy set for selecting sentences based on their features [13] . The feature extraction techniques are used to obtain the important sentences in the text. Some of features are used in this research such as sentence length. Some sentences are short or some sentences are long. What is clear is that some of the attributes have more importance and some have less, so they should have balance weight in computations and we use fuzzy logic to solve this problem by defining the membership functions for each feature. Therefore, the features score of each sentence that we described in the previous step are used to obtain the significant sentences. In this proposed system, we use method to extract the important sentences from multiple documents using Fuzzy logic system.

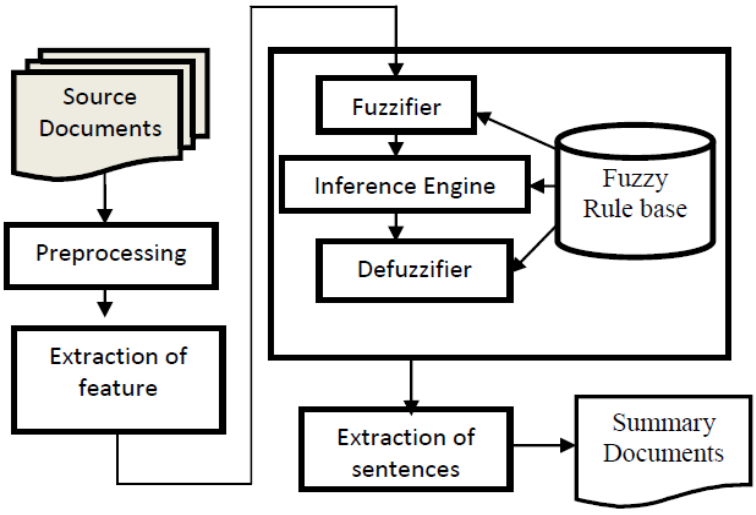

Figure 4.2 Fuzzy Logic System architecture.

4.3.1 Hindi Text Summarization Based on Fuzzy Logic

Fuzzy logic system design usually implicates selecting fuzzy rules and membership function. The selection of fuzzy rules and membership functions directly affect the performance of the fuzzy logic system. The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier, crisp inputs are translated into linguistic values using a membership function to be used to the input linguistic variables. After fuzzification, the inference engine refers to the rule base containing fuzzy IF -

THEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score.

The input membership function for each feature is divided into five fuzzy set which are composed of unimportant values (low (L) and very low (VL), Median (M) and important values (high (H) and very high (VH). The generalized Triangular membership function depends on three parameters a, b, and c as given by (12). A value from zero to one is obtained for each sentence in the output based on sentence features and the available rules in the knowledge base. The obtained value in the output determines the degree of importance of the sentence in the final summary.

$$f(x,a,b,c)=\max\left(\min\left(\frac{x-a}{b-a},\frac{a-x}{a-b}\right),0\right)$$

The parameters *a* and *c* set the left and right "feet" or base points, of the triangle. The parameter *b* sets the location of the triangle peak. In inference engine, the most important part in this procedure is the definition of fuzzy IF-THEN rules. The Important sentences are extracted from these rules according to our features criteria. Sample of IF-THEN rules shows as the following rule.

IF (NoWordInTitle is VH) and (SentenceLength is H) and (TermFreq is VH) and (SentencePosition is H) and (SentenceSimilarity is VH) and (NoProperNoun is H) and (NoThematicWord is VH) and (NumbericalData is H) THEN (Sentence is important)

Likewise, the last step in fuzzy logic system is the defuzzification. We used the output membership function. This is divided into three membership functions: Output {Unimportant, Average, and Important} to convert the fuzzy results from the inference engine into a crisp output for the final score of each sentence.

V. CONCLUSION

This paper discusses a multiple Hindi text Document 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. The proposed system use total 11 feature for calculating the sentences score. This summarizations system which is based on Fuzzy logic to improve the Quality of summary. The proposed method is implemented in Java & is under development.

REFERENCES

[1] Chetana Thaokar, Latesh Malik, "Test Model for Summarizing Hindi Text Using Extraction Method", IEEE Conference on ICT 2013.

[2] k.Vimal Kumar, Divakar Yadav, "An Improvised Extractive Approach to Hindi Text Summarization", Springer India 2015.

[3] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal of Emerging technology in web intelligence, VOL 2, NO 3, AUG 2010.

[4] Patil Pallavi, 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.

[5] Pallavi Patil, N.J.Kulkarni, "Text summarization using fuzzy Logic", International Journal of Innovative Research in Advanced Engineering,(IJIRAE),Vol 1, Issue 3,May 2014.

[6] Ms S.A.Babar, S. A. Thorat, "Improving Text Summarization using Fuzzy Logic & latent Semantic Analysis", International Journal of Innovative Research in Advanced Engineering,(IJIRAE), Vol 1, Issue 4, May 2014.

[7] Upendra Mishra, Chandra Prakash, "MAULIK:An Effective Stemmer for Hindi Language"

[8] Vishal Gupta, "Hindi Rule Based Stemmer for Nouns", International Journal of Advance Research in Computer Science and Software Engineering IJARCSSE Volume 4, Issue 1, Jan 2014.

[9] Priyanka Sarraf , Yogesh Kumar Meena, "Summarization of Document using Java", International Journal of Engineering Research & Technology (IJERT) Vol 3, Issue 2, February 2014.

[10] Vishal Gupta, Gurpreet Singh Lehal, "Features Selection and Weight learning for Punjabi Text Summarization", International Journal of Engineering Trends and Technology, Vol 2, Issue 2,2011

[11] Manish Prabhakar, Nidhi Chandra, "Automatic Text Summarization based on Pragmatic Analysis", International Journal of Scientific and Research Publications, Vol 2, Issue 5, May 2012.

[12] Hongling Wang Guodong Zhou, "Topic driven Multi-documents Summarization", IEEE International conference on Asian Language Processing, 2010.

[13] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security (IJCSIS), Vol. 2, No. 1, 2009.

[14] Arman Kiani, M. R. Akbarzadeh, "Automatic Text Summarization Using Hybrid Fuzzy GA-GP", IEEE International Conference on Fuzzy Systems, July 16-21, 2006.