Academic Science

International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 4, Special Issue
March 2015

A Comparative Study Of Hindi Text Summarization Techniques: Genetic Algorithm And Neural Network

Deepali P kadam
PG Student
Department of CSE
Datta Meghe COE,
Airoli,Navi Mumbai,India

Mrs. Nita Patil
Professor
Department of CSE
Datta Meghe COE
Airoli,Navi Mumbai,India

Mrs. Archana Gulathi
Professor
Department of CSE
Datta Meghe COE
Airoli,Navi Mumbai,India

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. As it helps in information retrieval, it also may lead you towards remedy for Information overload issue in today's world. In our approach, we have selected sentence extraction method for creating a summary. Sentences to select depend on the scores gained by the sentences. Higher the scores 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. We propose the evaluation of an automatic text summarization approach based on sentence extraction using genetic algorithm and neural network to come up with better quality result. Hindi is taken as a study language for the proposed work.

Keywords: Hindi text summarization, Genetic algorithm, feature extraction, Neural Network.

I. INTRODUCTION

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, etc, for readers, it is always convenient to read point to point specific gist of the document. 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.

Summarization can be of two types: Extractive and Abstractive. In our proposed system, we have chosen extractive summarization for the study purpose. What characteristics a sentence should possess to grab the position in the summary, is the core question to be answered. These characteristics are called as features and extraction of these features calculates the overall score a sentence would weigh. In our system, we have suggested six statistical and two linguistic features to be extracted. We are proposing two machine learning techniques Genetic Algorithm (GA) and Artificial Neural Network (ANN) for the sentence extraction and ranking. It is then followed by the comparative study of both the algorithms.

We have considered Hindi as a language of Study. It is written in the Devanagari script which has largest alphabet set. Hindi is an official language of India. It the native language of most people living in Delhi, Chhattisgarh, Himachal Pradesh, Chandigarh, Bihar, Jharkhand, Madhya Pradesh, Haryana, and Rajasthan. So for people who do not know English but want to read articles on the Internet, automatic summarization would play lion's role in it. While performing related search, it is observed that a lot of work has been done on English language as ample amount of resources are readily available for the same. Relatively very few have shown interest in the case of Hindi language. It motivated us for considering Hindi as a study language.

II. LITERATURE SURVEY

TABLE I
RELATED EXTRACTIVE SUMMARIZATION SEARCH

Name	Year	Method	Features
Automatic Text Summarization Using: Hybrid Fuzzy GA-GP [12]	2006	Hybrid Fuzzy GA-GP	Title feature, Sentence position, sentence length, no. of thematic words, no. of emphasize words
Bengali Text Summarizati-on by Sentence Extraction [5]		Extraction Method	Thematic term, positional value, sentence length,
Extractive Sentence Segments for Text Summarization : A Machine Learning Approach [3]	2000	Decision tree algorithm, Naïve Bayesian classifier, inter pattern distance based constructive neural network learning algorithm	Paragraph number, no. of bonus words, TF etc
A language independent approach to multilingual text summarization [10]			Title Feature, Thematic term, title feature
A hybrid approach to automatic text summarization [9]		KCS (K-mixture connective-strength)	TF-IDF
Fuzzy Logic Based Method for Improving Text Summarization [7]	2009	Fuzzy Logic	Title feature, Sentence Length, Term Weight, Sentence Position, Sentence to Sentence Similarity, Numerical Data
Automatic Text Summarization with Neural Networks [11]	2004	Neural Network	Sentence length, title words, thematic words, Sentence location in paragraph

Fuzzy Genetic Semantic Based Text Summarization [2]		Fuzzy Logic, Genetic Algorithm	Title feature, Sentence length, Term weight, Proper noun, Thematic word, Numerical data

III. TEXT SUMMARIZATION AND EXTRACTION TECHNIQUES

A. Definition

Text summarization is the process of distilling the most important information from the set of sources to produce an abridged version [1].

B. Types of Text Summarization

Text summarization can be performed in two different approaches: extraction and abstraction.

1)Extraction: 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: 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.

C. Types of Extraction Method

Extraction method is further classified as: Statistical, Linguistic and Hybrid approach.

1) 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.

2) Linguistic Method: In this, method needs to be aware of and know deeply the linguistic knowledge, so that the computer will be able to analyze 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.

3) Hybrid Method: It optimizes best of both the previous method for meaningful and short summary.

IV. PROPOSED SYSTEM

Flow of our proposed system is as follows:

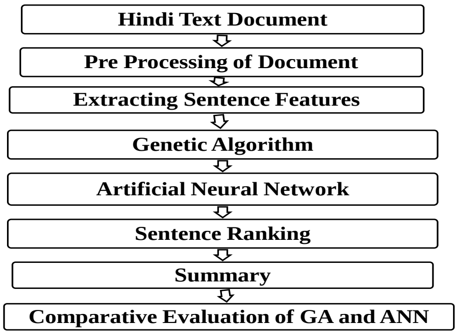

Hindi Text Document
Pre Processing of Document
Extracting Sentence Features
Genetic Algorithm
Artificial Neural Network
Sentence Ranking
Summary
Comparative Evaluation of GA and ANN

Fig . 1 : Flow of proposed system

Extractive text summarization process can be divided into two steps: Pre Processing step and Processing step.

A. Preprocessing 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. It includes:

1) 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 [3]. In Hindi, sentence is segmented by identifying boundary of sentence that ends with purnaviram(|).

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 (विराम चिन्ह) in Hindi language consists of पूर्ण विराम (|), उपविराम (:), अर्ध विराम (;), etc.

3) Stop Word Removal : Stop Words include function words, articles, prepositions, conjunctions, prefix, postfix, etc. i.e. common words that carry less important meaning than keywords.For example:

TABLE II
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
सर्वन ाम	आप, तू ,यह , वह ,कुछ		

They have no semantic as such and do not aggregate relevant information to the task. Also they make the text look heavier and are insignificant. Hence should be eliminated.

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.

B. 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 to1. Six statistical and two linguistic features are used as follows:

1) Average TF-ISF (Term Frequency Inverse Sentence Frequency): TF-ISF stands for term frequency-inverse document frequency and the tf-isf weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. 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).

$$TF = \frac{\text{Word(term)occurence in sentence(Si)}}{\text{Total no.of words in Sentence(Si)}}$$

$$ISF = \log\left(\frac{\text{Total no. of sentences}}{\text{No. of semmtences containing the term}}\right)$$

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:

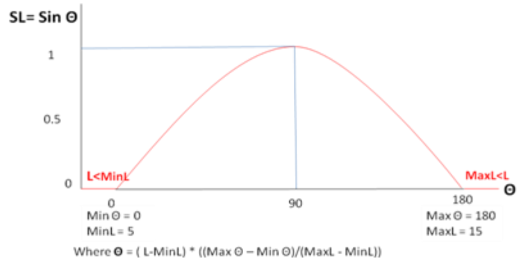
$$\text{AVG TFISF}(S_I) = \sum \text{TF} * \text{ISF}$$

2) Sentence Length: 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. Consider L = Length of Sentence

MinL = Minimum Length of Sentence (= 5 in our experiment)

MaxL = Maximum Length of Sentence (=15 in our experiment)

Min Θ = Minimum Angle (0) and Max Θ = Maximum Angle (180)



SL = 0 if L < MinL or L > MaxL
Otherwise
SL = Sin ((L-MinL) * ((Max Θ – Min Θ)/(MaxL - MinL)))

Fig. 2 : Sentence Length

3) Numerical Data: 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.

$$\text{ND} = \frac{\text{No. of Numerical Data in a Sentence}}{\text{Sentence Length}}$$

4) Sentence Position: 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. For remaining sentences: SP = Cos ((CP - Min V)*((Max Θ - Min Θ) /(MaxV - Min V))) where

TRSH = Threshold Value (10% in our exp.)

MinV = NS * TRSH (Minimum Value of Sentence)

MaxV=NS*(1-TRSH)(MaximumValueof Sentence)

NS = Number of sentences in document (50 in our exp.)

Min Θ = Minimum Angle (0) and Max Θ = Maximum Angle (360)

CP = Current Position of sentence [6]

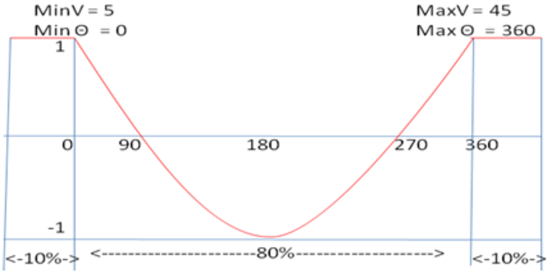


Fig. 3 : Sentence Position

But the values which are to be used as an input for GA or ANN have to lie between 0 and 1. So we need to apply normalization for above calculated values, as follows:

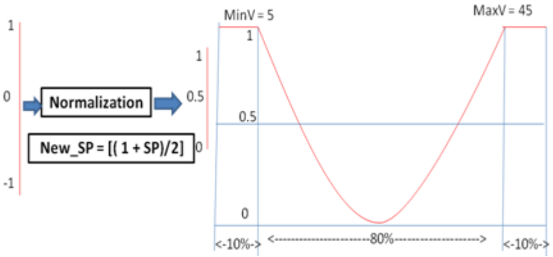


Fig. 4 : Sentence Position after Normalization

5) Proper Noun Feature: Proper noun is name of a person, place and concept etc. The sentence that contains more proper nouns (name entity) is an important and hence its probability to be present in a summary also increases. The score for this feature is :

$$\mathrm{PN} = \frac{\text{No. of Proper nouns in S}}{\text{Sentence length of S}}$$

6) Thematic Word Feature: The most frequent content words are defined as thematic words. A small number of thematic words is selected and each sentence is scored as a function of frequency [9]. Terms that occur frequently in a document are probably related to its topic. Therefore, we expect a high Occurrence of thematic words in salient sentences.

$$\mathrm{TW} = \frac{\text{No. of Thematic Words in S}}{\text{Max(No. of Thematic Words)}}$$

7) Sentence to Sentence Similarity: For each sentence s compute the similarity between s and each other sentence 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. We have chosen the following approach formulated in [6].

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

$$\mathrm{Sim}(i,j) = \frac{\text{Number of words occurred in Sentences(Sj)}}{\text{WT}}$$

where, N = Number of Sentences and W_T = Total Words in Sentence Si

8) Title Word Feature: Title itself is a kind of smallest summary that represents the gist or theme of the document. Obviously the words in the title carry higher weight and make the sentences containing them a possible candidate to be included in a summary. The weight is the count of the common words between a sentence and the title.

$$\mathbf{TS} = \frac{|\text{(words in a Sententence)} \cap \text{(words in a Title)}|}{|\text{Total words in Title}|}$$

C. The Methods

The goal of text summarization based on extraction approach is sentence extraction. The features score of each sentence that we described in the previous section are used to obtain the significantly important sentences. We propose to use two methods: text summarization based on Genetic Algorithm (GA) and Artificial Neural Network (ANN).

1) Genetic Algorithm: Genetic algorithm (GA) replicates the theme of Darwinian evolution and also the concept of the survival of the fittest. A purpose of this global search technique is to optimize a set of parameters. The fundamental steps consist of operations:

a. Initial Population : GA starts with the random generation of an initial set of individuals or chromosomes, the initial population. The genetic information is encoded in an array bit string [0, 1] of fixed length, called the parameter string or individual. Each individual represents a possible solution to the examined problem. This gene pool is formed by random function.

b. Fitness Function : The fitness value reflects how good or fit a chromosome is compared to the other chromosomes in the population. Better the chromosomes, higher the chance of survival and reproduction that can represent the next generation! Bad ones with worst fitness value are usually discarded. We define fitness function as

$\mathrm{F(S)} = \sum_{j=1}^{8} \mathrm{fj}(s)$

fj(s) is feature value of sentence (8 features, so It varies from 1 to 8)

3. c. Selection : With the reference of values provided by fitness function, chromosomes are selected from the population of parents to mate during reproduction and are kept in a mating pool. We use Roulette wheel selection method, where there is a chance that some weaker solutions may survive the selection process; this

is an advantage; as they may include some component which could prove useful following the recombination process.

d. Genetic operators: Crossover and Mutation : Cross over operator is applied to the mating pool which is enriched with better individuals with a hope that it would create better strings. It makes offspring of good strings but does not create new ones. During mutation, bits are flipped at random. Mutation may be applied to offspring produced by crossover or, as an independent operator, at random to any individual in the population. These processes will continue until the fitness value of individuals in the population converges or fixed number of generations reached i.e. until the system ceases to improve.

e. Sentence Ranking: The document sentences are scored. Using GA, best chromosome is selected after the specific number of generations. Then using Euclidean distance formula distance between sentence score and the fittest chromosome is evaluated. Sentences are sorted based on the ascending order of distance values. Depending on the compression rate sentences are extracted from the document to generate summary. [6]

2) Artificial Neural Network: An Artificial Neural Network (ANN) has been proved to be a very powerful tool for information processing paradigm. In our case, we have utilized the feed forward back propagation method. The way the network work will be:

a. One has to present a training sample (Hindi dataset in our case) to the neural network.

b. Then it will compare the network's output to the target output from that sample dataset. Desired output will be '1' if sentence is meant to be considered for the summary and it will be '0' if not.

c. Calculate the error in each output.

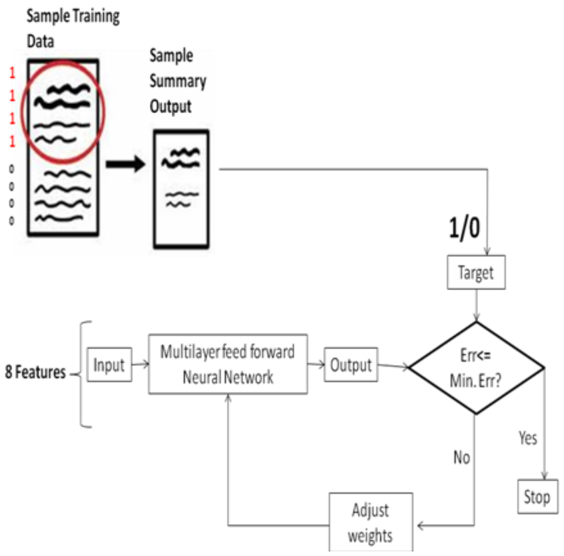


Fig. 5 : Neural Network Process

So, with the features and target output, we have a dataset which we can use to train the network. Network observes the pattern of the dataset and gets trained. For each neuron, the network calculates what the target output is and a scaling factor, how much lower or higher the output must be adjusted to match the target output. Network performs better if it is trained more.

Usually, networks for back propagation methods are multi layered. Here our basic network contains 8 input and one output layer. Though for performance measurement, we might modify our network frequently.

The results of these methods would be compared by any of the evaluation methods suitable for both of them. Available evaluation methods are ROUGE (Recall-Oriented Understudy for Gisting Evaluation), intrinsic methods that attempt to measure summary trustfulness by human experts, precision methods and recall methods.

V. CONCLUSION

We have proposed a comparison between two machine learning models for text summarization. These models are Genetic Algorithm and Neural Network namely. We have considered text summarization which is based on sentence extraction method. In the flow of proposed approach first feature extraction comes, then sentence scoring and lastly selection of higher ranked sentences as a summary. Six statistical and two linguistic features are used for this single document summarization. We have considered the Hindi, an official language of India, as a language of study. The proposed system is under development.

REFERENCES

[1] Inderjeet Mani, Therese Firmin, Beth Sundheim, "The TIPSTER SUMMAC Text Summarization Evaluation", EACL, 1999

[2] Ladda Suanmali1, Naomie Salim2 and Mohammed Salem Binwahlan3, " Fuzzy Genetic Semantic Based Text Summarization", Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011

[3] Wesley T. Chuang and Jihoon Yang, "Extractive Sentence Segments for Text Summarization : A Machine Learning Approach", ACM, 2000

[4] Duncan Temple Lang, "Word Stemming in R", 2004

[5] Kamal Sarkar, "Bengali Text Summarization By Sentence Extraction"

[6] Chetana Thaokar[1],Dr.Latesh Malik[2], "Test Model for Summarizing Hindi Text using Extraction Method",IEEE Conference on Information and Communication Technologies, 2013

[7] Ladda Suanmali1,Naomie Salim2 and Mohammed Salem Binwahlan3, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009

[8] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, VOL. 2, NO. 3, August 2010

[9] Te-Min Chang,Wen-Feng Hsiao, "A hybrid approach to automatic text summarization", IEEE, CIT, 2008

[10] Alkesh Patel, Tanveer Siddiqui, U. S. Tiwary, "A language independent approach tomultilingual text summarization" ,Conference RIAO,2007

[11] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", SECOND IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS, JUNE 2004

[12] Armaan Kiani, M. R. Akbarzadeh, "Automatic Text Summarization Using : Hybrid Fuzzy GA-GP", IEEE, July 2006