**Text Summarisation in Hindi**

A Final Year Project report

*Prepared by*

Abhishek Kumar (IT – 10114008)

Nishant Kumar (IT -10114017)

Of

Indian Institute Of Information Technology, Kalyani

West Bengal

( B.Tech Batch 2014-18 )

**Under the guidance of**

Dr. Manjira Sinha

**HINDI TEXT SUMMERISATION**

# **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 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

# 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 one machine learning techniques Genetic Algorithm (GA) for the sentence extraction and ranking. It is then followed by the comparative study of 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 important 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** | **METHODS** | **FEATURES** |
| Automatic Text Summarization Using: Hybrid Fuzzy GA-GP | 2006 | Hybrid Fuzzy GA-GP | Title feature, Sentence position, sentence length, no. of thematic words, no. of emphasize words |
| Bengali Text Summarization-on by Sentence Extraction | 2003 | Extraction Method | Thematic term, positional value, sentence length |
| Extractive Sentence Segments for Text Summarization : A Machine Learning Approach | 2000 | Decision tree algorithm, Naïve Bayesian classifier, inter pattern distance based constructive neural network learning algorithm | Paragraph number, no. of bonus words, TF etc. |
| Fuzzy Logic Based Method for Improving Text Summarization | 2009 | Fuzzy Logic | Title feature, Sentence Length, Term Weight, Sentence Position, Sentence to Sentence Similarity, Numerical Data |
| Automatic Text Summarization with Neural Networks | 2004 | Neural Network | Sentence length, title words, thematic words, Sentence location in paragraph |

# 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.

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 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.
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:

## 1 : Flow of proposed system

Extractive text summarization process can be divided into two step

Pre Processing step and Processing step.

### A. Preprocessing Step

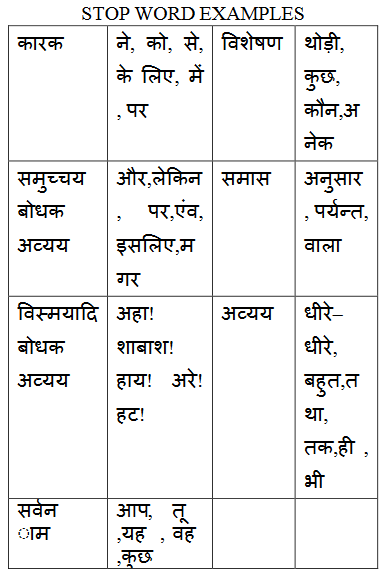
We need to prepare data for further processing. This intermediate preparation stage is called a Preprocessing step which is a structured representation of the original text. It includes:

1) 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( **|** ).

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 (**:**), अर्ण 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.



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 (t) = (Number of times term t appears in a document) / (Total number of terms in the document).**

**ISF (t) = log e (Total number of documents / Number of documents with term t in it).**

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:

**Avg TFISF (St) = ∑ TF\*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)))**

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.

**ND= no of Numerical Data in Sentence/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 - Min ) /(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 = Minimum Angle (0) and**

**Max = Maximum Angle (360)**

**CP = Current Position of sentence**

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 :

**PN= no of Proper nouns in S/ Sentence length of S**

6) Sentence to Sentence Similarity: For each sentence s compute the similarity between s by creating a weighted graph matrix 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.

**SS = and**

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

## 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.

Sentence Ranking: The document sentences are scored. 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.

## V. CONCLUSION

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. Five statistical and one linguistic features are used for this single document summarization. We have considered the Hindi, an official language of India, as a language of study.

## 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”