Feature Extraction in Hindi Text Summarization

Gunjan Pareek, Deepa Modi
Department of Computer Science & Engineering
Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur
Email: gunjanpareek1611@gmail.com
Received 19 August 2016, received in revised form 24 August 2016, accepted 27 August 2016

Abstract: **Technology has flooded human being with information, we just need to dial the text and search a complete heap of data which is later segregated according to its usefulness. However, this segregation is very difficult and involves usage of a brilliant mind. Here comes the use of Automatic Text Summarization (ATS), which condenses this information into useful information, saving user's time and attracts a complete ground for research work in Natural Language Processing (NLP). In this paper various statistical and linguistic features for a sentence are discussed. Based upon these features, weight is assigned to every sentence. According to this weight, importance of a sentence is decided into the summary. The summary generated by this method covers maximum theme with less redundancy. This work is done for the Hindi language.**

Key Words: **Hindi text summarization, natural language processing, statistical and linguistic features, SOV qualification.**

1. INTRODUCTION

Text summarization is defined as a condensed overview of the source text into a shorter version, without affecting the information content of the original text. In the current scenario for reading large amounts of text, text summarization is an important technique. According to Eduard Hovy, "A summary can be defined as a text that is produced from one or more texts, that contains a significant portion of the information from the original text, and that is no longer than half of the original text" [1]. Text summarization is useful in everyday life like headlines of news, abstract summary of technical paper etc.

Extraction and Abstraction text summarization are the two categories of automatic text summarization. In extractive text summarization, sentences or phrases are selected with highest importance from the original one and new shorter text will be created by putting them together without making any change in the originality. In abstractive text summarization, linguistic methods are used to examine and interpret the text. The extraction method is widely used.

For this research work the Hindi language is taken into account. The Hindi language has been written in Devanagari script and contains the largest set of letters. Hindi is the official language of India. It is the native language of most of the states and people residing Fiji, Surinam, Mauritius and Nepal. There has been done so much work on various languages like English, Punjabi, and Bengali etc. These are the reasons to choose the Hindi language for this experiment.

The paper discusses the "importance of the sentences in the generated summary", is decided based on various statistical and linguistic features of sentences, so that meaningful sentences are extracted from the text. Based on these feature overall score is calculated for each and every sentence. In this work, we have suggested six statistical and two linguistic features to be applied. It uses Hindi WordNet for assigning appropriate part-of-speech to every word within a sentence for checking SOV of the sentence [2]. The paper is arranged in the following sections as, feature extraction methods are described in Section II. Related work is described in the Section III. The proposed procedure is mentioned in Section IV. Section V summarizes the paper and states the future work.

2. METHODS OF FEATURE EXTRACTION

Feature extraction methods are classified into three categories as described below,

2.1 Statistical Method

Statistical method deals with the statistical distribution of features like extraction of keywords, phrases etc. and this is done without comprehend the document. It uses classification and information retrieval technology. Classification classifies the sentences that can be part of the summary depending on the training of the data. While information retrieval technology uses various features like position, sentence or word appearance in a long document etc.

2.2 Linguistic Method

This method allows computer to analyze and choose suitable sentences and thus the linguistic knowledge is must to be known. It establishes term relationship in the document by grammar analysis, by tagging part-of-speech, usage of thesaurus and emerges out with valuable sentences. Parameters can be cue words, title feature, noun and verbs in the sentences [3].

2.3 Hybrid Method

For short and meaningful summaries both previous methods are combined in this approach.

3. RELATED RESEARCH

In the field of automatic text summarization a lot of work has been done for English like language but for Hindi language it is still a few. Surface and indicators was an important part of the decision of the earlier research used to summarize the text, but now a day's some advanced techniques are used. Some previous works are mentioned in *"Table I"*.

Table 1 : Related Work on ATS

Paper	Features for Extraction & Domain	Domain
H. Luhn, 1958[4]	The frequency of word and phrase	Technical Papers
H. Edmundson, 1969 [5]	Word frequency, cue phrases, title words and sentence location.	Articles
D. Das and A. Martins, 1995 [6]	Term and sentence weighting.	News
E. H. Hovy, 1998 [1]	Topic identification, interpretation and Generation.	News
H. Jing and K. McKeown, 2001 [7]	Lexical coherence, tf*idf score, cue phrases and Sentence positions.	Domain independent
S. Gholamrezazadeh, M.Salehi and B. gholamzadeh, 2004[8]	Name entity recognition or multiword.	News
A. Kiani-B and M. R. Akbarzadeh-T, 2006 [9]	Title and thematic words	News articles
L. Suanmali, N. Salim and M. S. Binwahlan, 2011[10]	Title Function, length, proper nouns and location of heavy items.	DUC 2002
C. Thaokar and L. Malik, 2013 [2]	Statistical and linguistic feature, SOV qualification.	Hindi text document

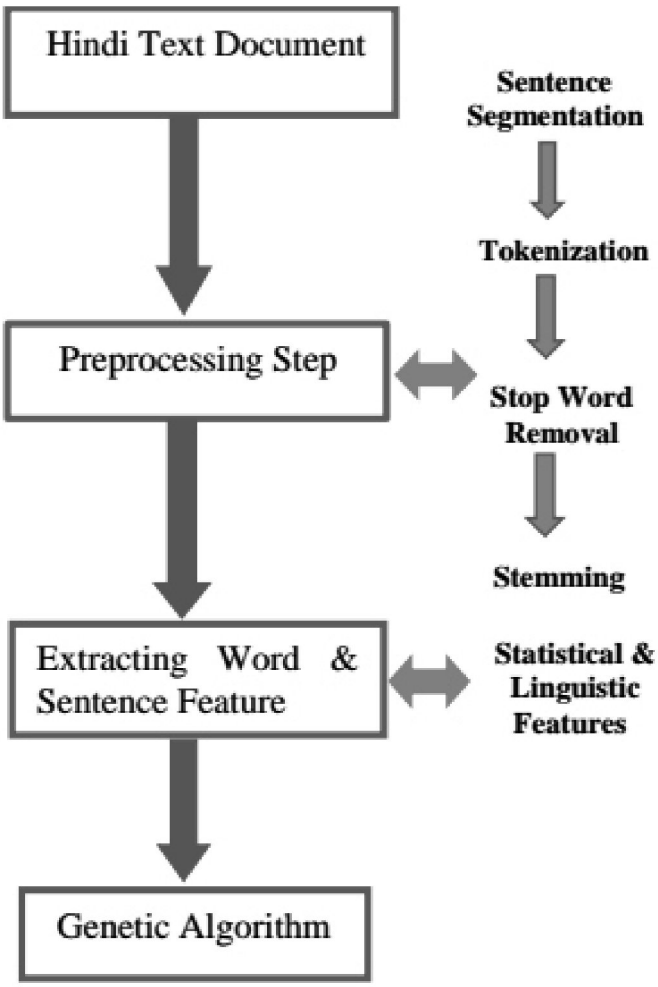


Fig.1. Proposed Model

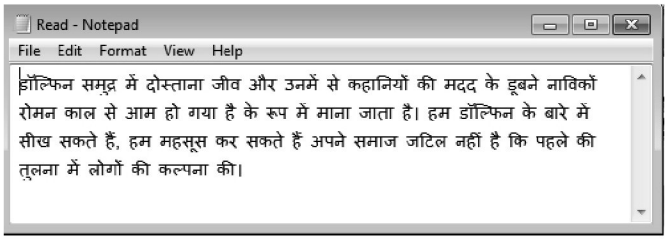
4. PROPOSED MODEL

Hindi text summarization includes preprocessing and feature extraction as the initial stage. Preprocessing is important as it

provides clean and adequate representation of source document, then extraction of feature terms is important as it helps in extracting most relevant sentences. As shown in *"Fig.1,"* two major steps, preprocessing and feature extraction are included in the very initial stage of our proposed model.

4.1 Preprocessing Step

It prepares the text document for further analysis. A sample Hindi text is shown in *"Fig. 2"*. Preprocessing involves three steps as defined below.



4.1.1 Sentence Segmentation

In this process, the text document is divided into constituent sentences. In Hindi, sentence is segment when sentence ends with purna viram (।) and Question mark (?). *"Fig. 3"* shows the result of sentence segmentation.

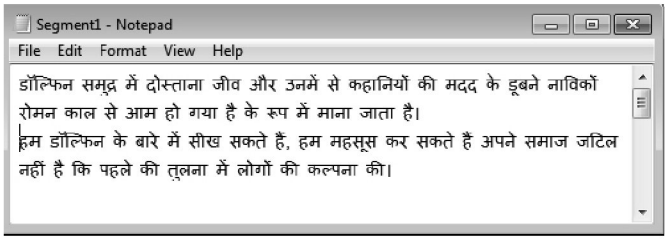


Fig.3. Result of sentence segmentation

4.1.2 Tokenization

The sentence is further split into tokens or words in this process and spaces, commas and special symbol were identified. The result of tokenization is stated in *"Fig. 4"*.

Fig.4. Result of tokenization

4.1.3 Stop Word Removal

Stop words are the common words that don't have their own meaning. To remove stop words from tokens, this step is performed. Some of the examples of the stop words are articles, functional words, prepositions, conjunctions, etc. These should be eliminated for efficient and short summary. A list of total 170 Hindi stop words is prepared and used in this step. Some examples of Hindi stop words are mentioned in *"Table II"*. "*Fig. 5"* is showing the result of stop word removal step.

Table 2 : Stopword List: Some Examples

अत	उन	और
अपना	उनका	कई
अपनी	उनकी	कर
अपने	उनके	करता
अभी	उनको	करते
अंदर	उन्हीं	करना
आदि	उन्हें	करने
आप	उन्हों	करें
इत्यादि	उस	कहते
इन	उसके	कहा
इनका	उसी	का
इन्हीं	उसे	काफ़ी
इन्हें	एक	कि
इन्हों	एवं	कितना

Fig.5. Result after stop word removal

4.1.4 Stemming

In this step roots or stem words are identified to get the common origin, so that suffixes are removed from the words to get common origin of the words. Syntactically similar words such as plurals, verbal variations, etc. are considered same. For e.g. कहानी, कहानियों, कहानियाँ are considered similar as they all are derived from a stem word कहानी. The result of this step is shown in *"Fig. 6"*.

Fig.6. Result of stemming.

4.2 Feature Extraction

The analysis of documents for the text summarization is initiate in this phase. In this phase every sentence is represented by a vector of feature terms. These terms are used to check both statistical and linguistical features. For ranking, each sentence is given a score, this score ranging between 0 to 1. For each and every sentence the value of eight features are calculated.

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

In this step, analysis of distribution of each word is performed over the document, and called term frequency (TF). Whereas, the term that are more useful than others are occur in few sentences, but maximum frequently within the sentence is called ISF.

TF = Word occurrence in the sentence (Si)/Total number of words in the sentence (Si) (1)

SF= log [Total Sentences in the document / Number of times the word occur] (2)

Then the average TF * IDF is calculated for each sentence and assign a weight accordingly.

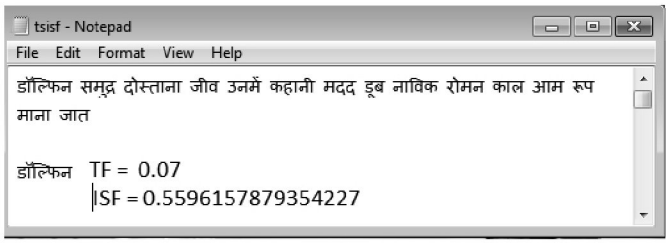


Fig.7. Result of TF-ISF feature

4.2.2 Sentence Length (F2)

This is a required feature in generating summary. In summary, short sentences such as names, date lines etc. are not important. Lengthy sentences may have lot of redundant data and hence are excluded in the summary. So we remove either too short or too long sentences from the summary.

SL = 0 if Len<Minimum _L or Len> Maximum_Len (3)

Otherwise,

SL = Sin ((Len –Minimum _L) * ((Maximum_θ–Minimum_θ) / (Max _L – Minimum_Len))) (4)

Where,

Len = Length of sentence

Minimum_L=the minimum length of the sentence (Minimum_L = 0 in this experiment)

Maximum_L=the maximum length of the sentence (Maximum_L = 25 in this experiment)

Minimum_θ = Minimum Angle (0^o)

Maximum_θ = Maximum Angle (180^o)

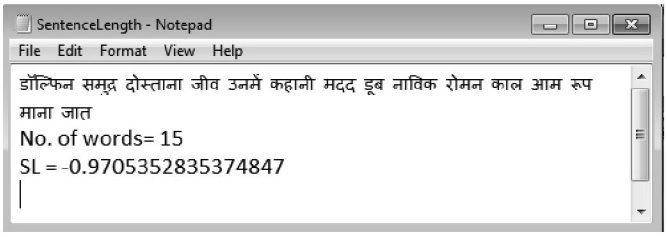


Fig.8. Result of sentence length feature

4.2.3 Numerical Data (F3)

It is used to represent important mathematical or statistical analysis providing vital information in a document.

Sentence = 1, if digit exist (5)

Otherwise,

Sentence = 0, if digit do not exist (6)

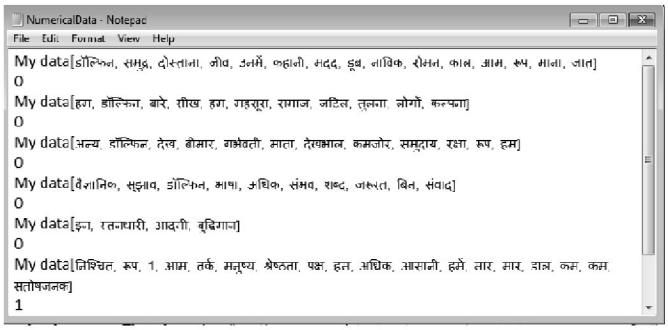


Fig.9. Result of numerical data feature

4.2.4 Sentence Position (F4)

The position of a sentence in the text is used to decide the importance of the sentence. As the theme of a document is defined in the starting and the conclusion or summarization is in the end of the text so a threshold value is taken to decide the number of sentences to be retained in the beginning and at the end.

S_P=1, for the sentence in the beginning and in the ending (7)

for Remaining sentences, F4 is calculated as

S_P=Cos((C_P–Minimum_V)*((Maximum_θ–Minimum_θ)/ (Maximum_V – Minimum _V))) (8)

Where,

Minimum_V = Nos*TRESH (Minimum Value of Sentence)

Maximum_V=Nos*(1 – TRESH) (Maximum Value of Sentence)

TRESH = Threshold value (10% in this work)

Nos = Number of sentences

Minimum_θ = Minimum Angle (0^o)

Maximum_θ = Maximum Angle (180^o)

C_P = the current position of the sentence

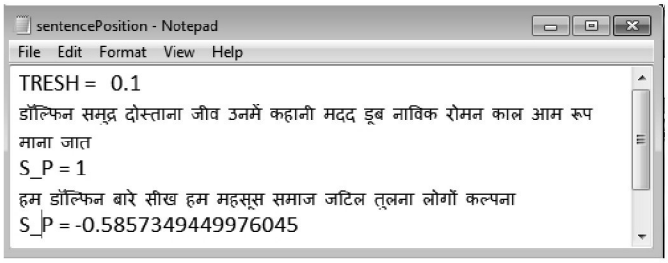


Fig.10. Result of sentence position feature

4.2.5 Title Feature (F5)

Title itself is smallest summary which represents the theme of the document. Words in the title carry higher weight and make the sentences including them a possible candidate to be used in the summary.

$$TF = \frac{WoS \cap WoT}{\text{Total words in title}} \quad (9)$$

Where,

WoS = Number of words in a sentence

WoT = Number of words in a title

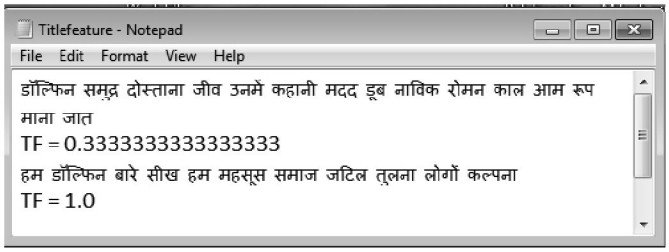


Fig.11. Result of the title feature

4.2.6 Sentence to Sentence Similarity (F6)

In order to determine the similarity between different sentences, sentence to sentence similar characteristic is use. For this match, stem words are used.

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

Sim (i, j) = WoS(Sj)/Total words in sentence(Si) (11)

Where,

NoS= Number of sentences

WoS = Number of words in a sentence (Sj)

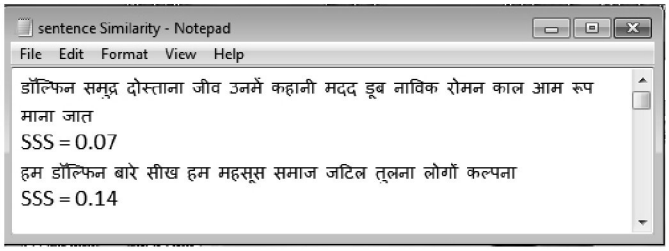


Fig.12. Result of sentence to sentence similarity feature

4.2.7 SOV Qualification (F7)

A sentence must include a subject and a verb. In Hindi, <Subject><Object><Verb> is a typical word order. This is why; the Hindi language is called "SOV" language. For SOV qualification each word of the sentence is marked with their respective part-of-speech tag [11-12]. The lexical database used in this study is Hindi WordNet 1.4, developed by IIT Bombay.

Every sentence in the text is examined for SOV qualification. If the first word in the sentence is a noun, then it is marked as the Subject_SOV of the sentence. This process continues till the end of the sentence, if the last word is a verb in the sentence then this sentence is qualified as SOV sentence. Only those sentences which are qualified as SOV will be used for further processing [2]. A sentence is checked for SOV qualification after removing the stop words. For example, डॉल्फिन समुद्र दोस्ताना

जीव कहानियों मदद डूबने नाविकों रोमन काल आम रूप माना जाता

Pos Tagging & Sov Qualification

Word	POS	SOV
डॉल्फिन	NOUN	Subject_SOV
समुद्र	NOUN	Object_SOV
दोस्ताना	ADJECTIVE	Object_SOV
जीव	NOUN	Object_SOV
कहानियों	NOUN	Object_SOV
मदद	NOUN	Object_SOV
डूबने	VERB	Object_SOV
नाविकों	NOUN	Object_SOV
रोमन	NOUN	Object_SOV
काल	NOUN	Object_SOV
आम	ADJECTIVE	Object_SOV
रूप	NOUN	Object_SOV
माना	VERB	Verb_SOV
जाता	VERB	Verb_SOV

4.2.8 Subject Similarity (F8)

The analysis of this step is similar to the previous step, as previous step is characterized by the subject of the sentence; in the same way in this step also we found whether the subject of the sentence is similar to the theme. Thematically similar function may be the same, in terms of the title and the sentence examination.

Sub_S = 1, if POS is noun and root value of title and sentence is equal (11)

Otherwise,

Sub_S= 0 (12)

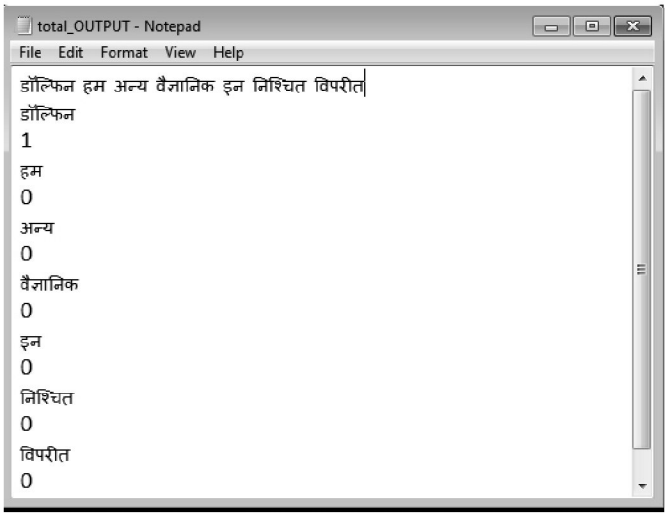


Fig.13. Result of subject similarity feature

5. CONCLUSION & FUTURE WORK

Hindi is our national language, and most spoken language in the India. In this work, an auto text summarization technique is proposed. This work explains six statistical and two linguistic features extraction method, to get significant sentences. This proposed method is already implemented in Java and would be used further to generate a summary and could be used by researchers to conclude and explain lengthy text in a very time efficient manner. In future, genetic algorithm would be used in this work to generate a summary of the Hindi text.

REFERENCES

[1] E. H. Hovy, "Automated Text Summarization," in *Ox-ford Handbook of Computational Linguistics*, Oxford University Press, 2005, pp. 583-598.

[2] C. Thaokar and L. Malik, "Test model for summarizing hindi text using extraction method," in *Information & Communication Technologies (ICT)*, 2013, pp. 1138-1143.

[3] P. Dehkordi, H. Khosravi and C. Kumar, "Text Summarization Based on Genetic programming," *International Journal of Computing and ICT Research,* vol. 3, no. 1, pp. 57-64, 2009.

[4] H. Luhn, "The automatic creation of literature abstract," *IBM journal of Research and development,* vol. 2, no. 2, pp. 159-165, 1958.

[5] H. Edmundson, "New methods in automatic extracting," *Journal of ACM (JACM),* vol. 16, no. 2, pp. 264-285, 1969.

[6] D. Das and A. Martins, "A survey on automatic text Summarization," *Literature survey for languages and statistics II course at CMU,* vol. 4, pp. 192-195, 2007.

[7] H. Jing and K. McKeown, "Cut and paste based text summarization," *Proceedings of the 1st North american chapter of the association for computational Linguistic Conference,* Stroudsburg, PA, USA, 2000, pp. 178-185.

[8] S. Gholamrezazadeh, M. Salehi and B. gholamzadeh, "A Comprehensive Survey on Text Summarization Systems," in *2nd International Conference on Computer Science and its Applications, Jeju, Korea (South),* 2009, pp. 1-6.

[9] A. Kiani-B and M. R. Akbarzadeh-T, "Automatic text summarization using hybrid fuzzy GA-GP," in *IEEE International Conference on Fuzzy System,* 2006, pp. 977-983.

[10] L. Suanmali, N. Salim and M. S. Binwahlan, "Fuzzy genetic semantic based text summarization. In Dependable, Autonomic and Secure Computing (DASC)," in *IEEE Ninth International Conference,* 2011, pp. 1184-1191.

[11] D. Modi and N. Nain, "Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method," *In Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing Springer India.,* India, 2016, pp. 241-247.

[12] M. Deepa, N. Maninder, N. Neeta and A. Mushtaq, "A Survey of Techniques for Two Level Corpus Annotation for Hindi," *International Bulletin of Mathematical Research*, vol 2, no. 1, pp. 194-206, 2015.

◆ ◆ ◆