A STATISTICAL MODEL FOR GIST GENERATION: A CASE STUDY ON HINDI NEWS ARTICLE

M. Varaprasad Rao[1], Dr B. Vishnu Vardhan[2], and P. Vijay Pal Reddy[3]

[1]Department of Computer Science, MIPGS, Hyderabad, AP, India
[2] Department of Computer Science & Engineering, JNTUH-J, Karimnagar, AP, India
[3] Department of Computer Science & Engineering, MREC,Hyderabad, AP, India

ABSTRACT

Every day, huge number of news articles are reported and disseminated on the Internet. By generating gist of an article, reader can go through the main topics instead of reading the whole article as it takes much time for reader to read the entire content of the article. An ideal system would understand the document and generate the appropriate theme(s) directly from the results of the understanding. In the absence of natural language understanding system, it is required to design an appropriate system. Gist generation is a difficult task because it requires both maximizing text content in short summary and maintains grammaticality of the text. In this paper we present a statistical approach to generate a gist of a Hindi news article. The experimental results are evaluated using the standard measures such as precision, recall and F1 measure for different statistical models and their combination on the article before pre-processing and after pre-processing.

KEYWORDS

Natural language understanding, precision, recall, F1 measure, sentence selection model, text model, informative word selection model, statistical model.

1. INTRODUCTION

In today the people are being exposed to more information available on the Internet, to select about what to read and what not to read or skip. The selection of information would be done based on two points; namely; Keywords and Bullet points. Based on this, a user can decide that whether a document to read, to skip or to scan. As explained in [1], gist is used to give the reader a sense (s) of what to expect from the following text in terms of form and content. Additionally, gist work as "intellectual and aesthetic controls [1]" by attracting the attention of the reader, and keeping them focused on the main theme(s).

It is unfortunate that automatic generation of themes has been a neglected issue in the field of text processing. Over the past decades, text processing studies have addressed many issues in text analysis, such as text categorization, automatic indexing, automatic abstracting, but automatic theme generation did not receive enough attention, even though it is strongly related to the above studies. The objective behind automatic indexing, automatic abstracting and automatic gist

DOI : 10.5121/ijdkp.2013.3502

generation are the same: identifying the main theme(s) and related topics in a specific document, or a document set. As defined in [2], text theme is "a specific subject that is discussed in some depth in a particular text, or in a number of related texts". In order to identify the text content, in this chapter statistical model was designed and experimented on Hindi text.

This paper is organized as follows: In Section 2 the research work related text summarization, short summary generation, gist generation are briefly explained. Section 3 describes the characteristics of Hindi language. Section 4 is dealt with the proposed statistical model. Section 5 is about data collection, evaluation measures and results analysis. The conclusions and further research are given in section 6. The sample document is given in section 7.

2. RELATED WORK

A lot of research in single document summarization has gone into finding out the relevant segments from the text, ranking them and finally generating the summary which expresses most of the important points. The task of gist generation is strongly connected to traditional text summarization [3] and emphasizes the extractive approach which selects words, sentences or paragraphs from the document to provide a summary.

Witbrock and Mittal interpreted the task of bullet point extraction as a twofold process. First, the system selects n words from the article that best reflect its content as content Selection process. Second, the best grammatical ordering of these n words is determined and termed it as surface realization. Each of these criteria is scored probabilistically, where the probability is estimated by prior collection of corpus statistics. Bayesian approach [4] [5], TF*IDF ranking [6], Text Model, Headline word position model [7] etc. have been used for determining the content selection probability. [8] presents another novel approach for selecting headline words. This chapter investigates the use of Singular Value Decomposition (SVD) as a means of determining if a word is a good candidate for inclusion in the gist. [9] describes a new approach towards informative word selection, viewing it as a variant of the Information Retrieval problem.

In [10] [11], the sentence extraction based method tries to pick the sentence which reflects the main content of the text. A number of researchers have looked into this problem of sentence extraction which carries the central idea of a corpus, as a solution of text summarization problem. The same techniques have been applied to the problem of gist extraction with a little variation [12]. [13] describes a procedure to automatically acquire topic signatures and evaluates the effectiveness of applying topic signatures to extract topic related sentences. More recently, some researchers have moved towards learning approaches [14], [15] that take advantage of training data. [16] shows how K Nearest Neighbor algorithm can be applied for gist generation. [15] describes another Machine learning (SVM) based summarization technique.

Another approach to construct gist by selecting words in order from the story, removes grammatical constituents from a parse of the lead sentence until a length threshold is met. This approach is called parse-and-Trim method for headline generation [17]. By treating topic generation as a variant of the Machine Translation problem, Kennedy and Hauptmann [18] came up with the generative approach using iterative Expectation Maximization algorithm. A number of researchers have attacked the problem of theme generation using NLP techniques. [19] [20] outline algorithms for computing lexical chains as an intermediate representation for automatic machine text summarization. A HMM [21] based summarization has been presented in [22] [23]. This idea has been extended for gist Generation for News Stories in [24].

3. LANGUAGE CHARACTERSTICS

Hindi is the national language of India. It is one of several languages spoken in different parts of the sub-continent. National should be understood as meaning the official or link language. The homeland of Hindi is in the North of India, but it is studied, taught, spoken and understood widely throughout the sub-continent, whether as mother tongue or as a second or a third language. Hindi is written in Devanagari script. The script is phonetic, so that Hindi, unlike English, is pronounced as it is written. Therefore, it is to learn the characters of the script and the sounds of the language at the same time.

There are 33 consonants and 11 vowels in Hindi. Additionally, there are also many conjunct consonants. Hindi consonants are divided into groups on the basis of phonetic properties of their formations such as plosives, nasals, fricatives, flapped and tapped sounds, and semi vowels. Each Devanagari script character represents a syllable, not the alphabet. English preposition-like Hindi words or "postpositions" follow their related words. Hindi nouns change from direct to oblique grammatical case when followed by postpositions. Oblique noun qualifying adjectives also acquire oblique usage. Hindi sentence syntax follows subject-object-verb structure.

4. PROPOSED MODEL FOR ARTICLE GIST GENERATION

The statistical model proposed in this chapter contains four phases such as pre-processing, statistical model, informative word selection and clustering as shown in Figure 4.1. The article was pre-processed by doing normalization, stop words removal and stemming. Then the words in the pre-processed article are scored using the combination three individual models. Top nine scoring words are selected as informative words and then these words are mapped onto the original article at their first occurrence. The Key phrases are formed using trigram clustering on the original article. The different models are explained below.

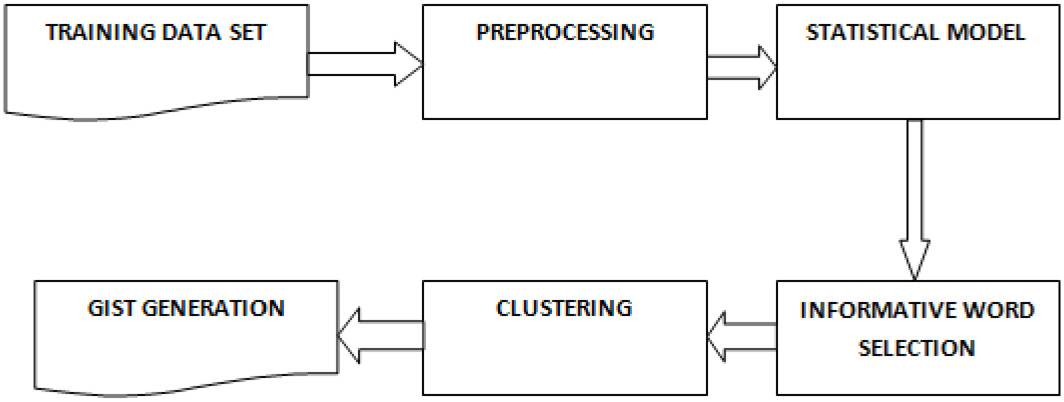


Figure 4.1: Statistical model for Hindi Article Gist Generation

4.1 Sentence Position Model:

Sentence position information has long proven useful in identifying topics of texts. This idea is applied to the selection of informative words. Given a sentence with its position in text, the likelihood that it would contain the first appearance of an informative word in the key phrase calculated as follows:

$$CountPos_i = \sum_{k=1}^{M} \sum_{j=1}^{N} P(G_k / W_j)$$

$$P(G / Pos_i) = \frac{CountPos}{\sum_{i=1}^{P} CountPos_P}$$

For each sentence position i over all M texts in the collection and over all the words in the M key phrases (each containing up to N words), CountPos records the number of times where sentence position i has the first appearance of any informative word. $P(G_k|W_j)$ is a binary feature. This is calculated for all sentence positions from 1 to P. Resulting $P(G|Pos_i)$ represents each sentence position containing one or more informative words.

4.2 Informative Word Position Model:

For each content word W_g, it would most likely first appear at sentence position

Pos_i:

$$P(Pos_i / W_g) = \frac{Count\,(Pos_i, W_g)}{\sum_{i=1}^{P} Count\,(Pos_P, W_g)}$$

In this model, information collected for each content word W_g.

4.3 Text Model:

This model captures the correlation between words in text and words in key phrases:

$$P\left(G_w / T_w\right) = \frac{\sum_{j=1}^{M} \left(docTf\ (w, j) * titleTf\ (w, j)\right)}{\sum_{j=1}^{M} docTf\ (w, j)}$$

Where

- docTf(w,j) is the term frequency of word ‘w’ in the jth document of all M documents in the collection.
- titleTf(w,j) is the term frequency of word w in the jth title.

- Gw and Tw are words that appear in both the theme and body of the text. For each instance of Gw and Tw pair, Gw=Tw.

The combination of sentence position and text model is calculated as follows:

$$P(G/W_i) = P(G/Pos_i) * P(Pos_i/W_g) * P(Gw_i/Tw_i)$$

5. Empirical Evaluations

5.1 Test Collections

The experimental dataset was gathered from various Hindi news chapters from the web during the year 2011 – 2011. There are a total of 1000 documents and corresponding gists in the corpus. The evaluation was based on the cumulative unigram overlap between the n top-scoring words and the reference headlines.

5.2 Evaluation Methods

The experimental results are evaluated using the precision, recall and F1 measures to compare the machine identified informative words with the human assigned content words. The above matrices have been proved as good evaluation matrices in the field of information retrieval.

The F1 measure can be calculated by using precision and recall as in following equation.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

where, precision is the number of common words among machine identified informative words ($G_{machine}$) and human assigned content words (G_{human}) divided by the number of machine identified informative words as in following equation:

$$precision = \frac{G_{machine} \wedge G_{human}}{G_{machine}}$$

recall is defined as the number of common words between $G_{machine}$ and G_{human} and divided by the number of words in the human assigned content words as in following equation:

$$recall = \frac{G_{machine} \wedge G_{human}}{G_{human}}$$

Precision shows the percentage of words being correctly identified by the machine with respect of the human generated gist. Where as recall gives the percentage of correct words that computer has selected, among the gist assigned by human subjects. F1 measure balances both precision and recall measures. The First highest scored nine words were selected as informative words, as it is the average number of content words in the corpus.

5.3 Results and Discussion

Table 4.1: F1 Measures for Eight Possible Combinations

F1 Measure	Sentence position model	Informative word selection model	Text model	Combination of all models
Before preprocessing	0.253	0.196	0.238	0.294
After preprocessing	0.514	0.352	0.486	0.602

The data corpus is evaluated using F1measure for eight possible combinations. Informative words are selected from the original article with sentence position model, informative word selection model and text model individually. Then F1 measure is calculated using precision and recall. Similarly the F1 measure calculated using the combination of all the models with equal weights. The results are specified in the Table 4.1. From the results in the influence of the stop words on informative words selection is notified. Similarity the influence of the three models independently and their combination is also observed.

The documents are pre-processed using different pre-processing phases and informative words are selected from the stemmed documents. The remaining four possible combinations are performed on the pre-processed document. The results are specified in the Table 4.1.

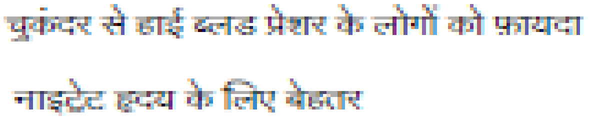
6. Conclusion And Future Scope

In this paper, we proposed a statistical model to select the informative words from the article. The selections of the informative words are based on combination sentence selection model, content word position model and text model. From the results we can conclude the influence of the stop words in the process of informative word selection and also the influence of the models individually and their combination for informative word selection. The generated gist from the machine was appropriate to the human generated gist. The results are encouraging.

There is a scope for further enhancement of the existing model for increasing the accuracy of the gist by generating the gist from the summary of the article instead of from the original article. It is also possible to enhance the accuracy using word net and ontology to combine the related words into a single word.

7. SAMPLE ARTICLE

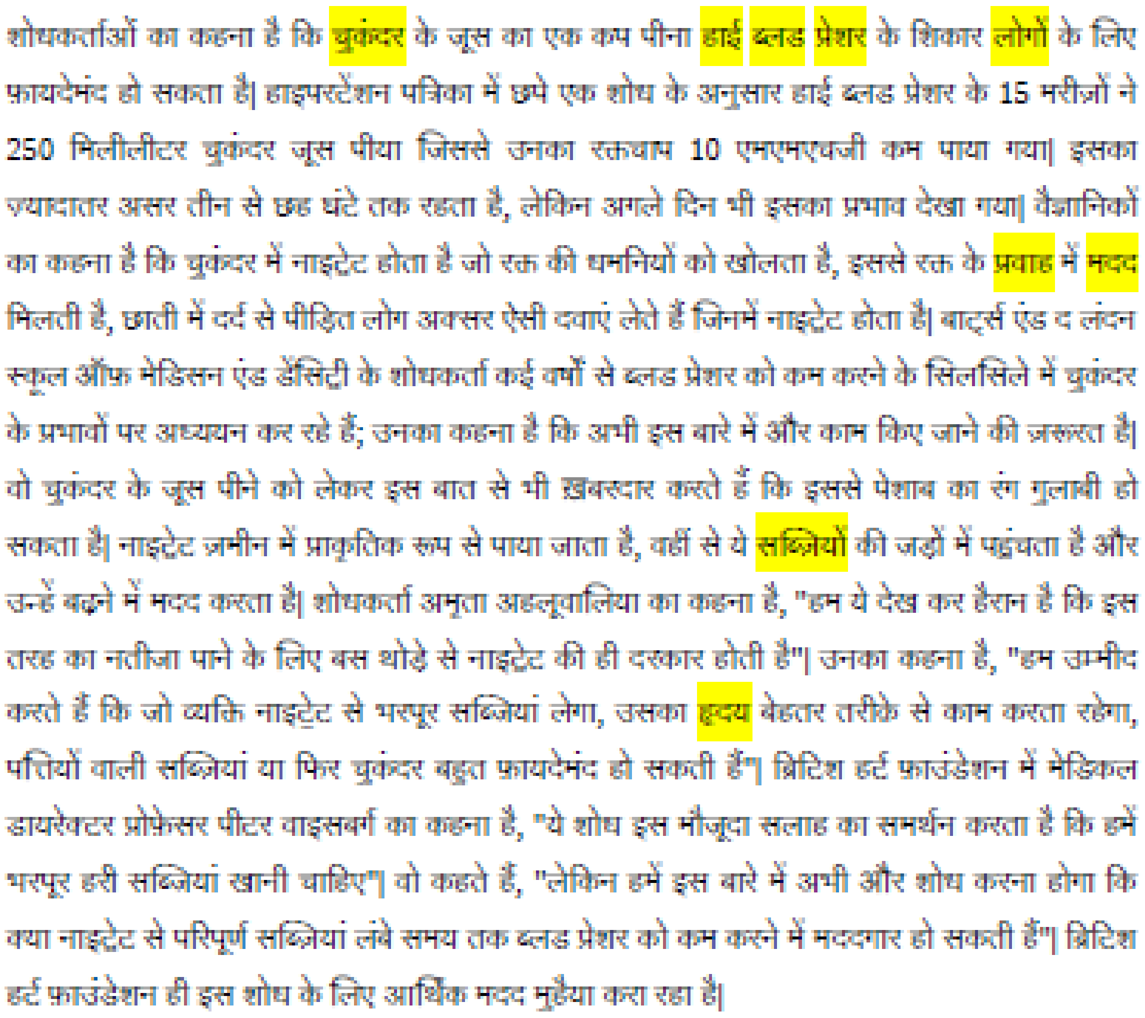
7.1 Human generated Gist



चुकंन्दर से हाई ब्लड प्रेशर के लोगों को फ़ायदा

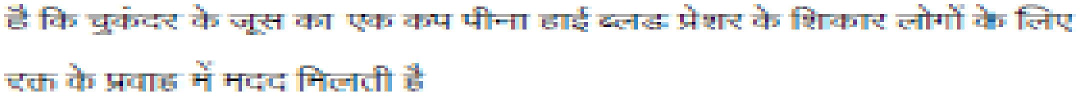
नाइट्रेट हृदय के लिए बेहतर

7.2 Article



शोधकर्ताओं का कहना है कि चुकंन्दर के जूस का एक कप पीना हाई ब्लड प्रेशर के शिकार लोगों के लिए फ़ायदेमंद हो सकता है| हाइपरटेंशन पत्रिका में छपे एक शोध के अनुसार हाई ब्लड प्रेशर के 15 मरीज़ों ने 250 मिलीलीटर चुकंन्दर जूस पीया जिससे उनका रक्तचाप 10 एमएमएचजी कम पाया गया| इसका ज्यादातर असर तीन से छह घंटे तक रहता है, लेकिन अगले दिन भी इसका प्रभाव देखा गया| वैज्ञानिकों का कहना है कि चुकंन्दर में नाइट्रेट होता है जो रक्त की धमनियों को खोलता है, इससे रक्त के प्रवाह में मदद मिलती है, छाती में दर्द से पीड़ित लोग अक्सर ऐसी दवाएं लेते हैं जिनमें नाइट्रेट होता है| बार्ट्स एंड द लंदन स्कूल ऑफ़ मेडिसन एंड डेंसिटी के शोधकर्ता कई वर्षों से ब्लड प्रेशर को कम करने के सिलसिले में चुकंन्दर के प्रभावों पर अध्ययन कर रहे हैं; उनका कहना है कि अभी इस बारे में और काम किए जाने की ज़रूरत है| वो चुकंन्दर के जूस पीने को लेकर इस बात से भी ख़बरदार करते हैं कि इससे पेशाब का रंग गुलाबी हो सकता है| नाइट्रेट ज़मीन में प्राकृतिक रूप से पाया जाता है, वहीं से ये सब्ज़ियों की जड़ों में पहुंचता है और उन्हें बढ़ने में मदद करता है| शोधकर्ता अमृता अहलूवालिया का कहना है, "हम ये देख कर हैरान है कि इस तरह का नतीजा पाने के लिए बस थोड़े से नाइट्रेट की ही दरकार होती है"| उनका कहना है, "हम उम्मीद करते हैं कि जो व्यक्ति नाइट्रेट से भरपूर सब्ज़ियां लेगा, उसका हृदय बेहतर तरीक़े से काम करता रहेगा, पत्तियों वाली सब्ज़ियां या फिर चुकंन्दर बहुत फ़ायदेमंद हो सकती हैं"| ब्रिटिश हर्ट फ़ाउंडेशन में मेडिकल डायरेक्टर प्रोफ़ेसर पीटर वाइसबर्ग का कहना है, "ये शोध इस मौजूदा सलाह का समर्थन करता है कि हमें भरपूर हरी सब्ज़ियां खानी चाहिए"| वो कहते हैं, "लेकिन हमें इस बारे में अभी और शोध करना होगा कि क्या नाइट्रेट से परिपूर्ण सब्ज़ियां लंबे समय तक ब्लड प्रेशर को कम करने में मददगार हो सकती हैं"| ब्रिटिश हर्ट फ़ाउंडेशन ही इस शोध के लिए आर्थिक मदद मुहैया करा रहा है|

7.3 Machine generated Gist



है कि चुकंन्दर के जूस का एक कप पीना हाई ब्लड प्रेशर के शिकार लोगों के लिए

रक्त के प्रवाह में मदद मिलती है

REFERENCES

[1] R.M. Person, “What’s in a title?’, Writer’s Digest, Vol. 65, Dec. 1985, pp. 13-14.

[2] G. Salton, J. Allan, C. Buckley, and A. Singhal, “Automatic analysis, theme generation, ans summarization of machine-readable texts”, Science, Vol. 264, Jun. 1994, pp. 1421-1426.

[3] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In SIGIR 99, Berkeley, CA, August 1999.

[4] M. Witbrock and V. Mittal. Ultra-summarization: A statistical approach to generating highly condensed nonextractive summaries. In SIGIR 99, Berkeley, CA, August 1999.

[5] M. Fuentes, M. Massot, H. Rodriguez, and L. Alonso. Mixed approach to headline extraction. In DUC, 2003.

[6] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In 4th ACM conference on Digital Libraries, pages 254–255, August 11- 14,1999.

[7] L. Zhou and E. Hovy. Headline summarization at isi. In DUC, 2003.

[8] S. Wan, M. Dras, C. Paris, and R. Dale. Using thematic information in statistical headline generation. In Workshop on Multilingual Summarization and Question Answering: Machine Learning and Beyond, ACL 03, July 2003.

[9] Rong Jin and Alexander G. Hauptmann. Learning to select good titlewords: An new approach based on reverse information retrieval. In 18th International Conference on Machine Learning (ICML), June 28-July 1, 2001.

[10] I. Demiros, H. Papageorgiou, and S. Piperidis. Sentence-based text summarization : Modelling and evaluation. In 2nd Hellenic Conf. On AI, SETN-2002, pages 103–114, 11-12 April 2002. [15] D. McDonald and H. Chen. Using sentence-selection heuristics to rank text segments in txtractor. In JCDL, pages 28–35, 2002.

[11] T.A.S. Pardo, L.H.M. Rino, and M.G.V Nunes. Gistsumm: A summarization tool based on a new extractive method. In 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken, pages 210–218, 2003.

[12] W. Kraaij, M. Spitters, and A. Hulth. Headline extraction based on a combination of uni- and multiocument summarization techniques. In Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics, July 2002.

[13] C. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In 18th International Conference on Computational Linguistics, COLING, 2000.

[14] W. T. Chuang and J. Yang. Extracting sentence segments for text summarization: A machine learning approach. In SIGIR, pages 152– 159, 2000.

[15] T. Hirao, K. Takeuchi, H. Isozaki, Y. Sasaki, and E. Maeda. Ntt/naists text summarization systems for tsc-2. In 3rd NTCIR Workshop, 2003.

[16] R. Jin and A. G. Hauptmann. Headline generation using a training corpus. In 2nd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING, 2000.

[17] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence, 139(1):91–107, 2002.

[18] P. Kennedy and A. G. Hauptmann. Automatic title generation for the informedia multimedia digital library. In ACM Digital Libraries, DL- 2000, May 2000.

[19] H. Gregory Silber and Kathleen McCoy. An efficient text summarizer using lexical chains. In 1st International Conference on Natural Language Generation, INLG, pages 268–271, June 2000.

[20] Barzilay, Regina, and M. Elhadad. Using kxical chains for text summarization. In Intelligent Scalable Text Summarization Workshop (ISTS97), 1997.

[21] L. R. Rabiner. A tutorial on hidden markov models and selected apllications in speech recognition. In A. Waibel and K. F. Lee, editors, Readings in Speech Recognition, pages 267–296, San Mateo, CA, 1990. Kaufmann.

[22] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary. Using hmm and logistic regression to generate extract summaries. In DUC, 2001.

[23] D. P. O'Leary J. M. Conroy. Text summarization via hidden markov models. In SIGIR, pages 406–407, 2001.

[24] D. Zajic, B. Dorr, and R. Schwartz. Automatic headline generation for newspaper stories. In Workshop on Automatic Summarization, July 2002.