Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Mixed Approach for Hindi News Article Gist Generation

M. Varaprasad Rao *
Department of CS, MIPGS, OU, Hyderabad, India

Dr. B. Vishnu Vardhan
Department of CSE, JNT UH-J Karimnagar, India

Abstract*— Huge news articles are reported and disseminated on the Internet. How to extract key information and save the reading time is the important issue. Gist is often used to summarize the content of the document. In this paper improved statistical approach based on keywords is presented. Gist generation problem can be viewed as finding informative keywords from the document and finding the proper way to combine these words to reflect a coherent and grammatical phrase. To improve the accuracy of the key phrase in reflecting the content, the summary of the document is first identified. The results show that proposed approach is more suitable for generating informative key phrases. This paper identifies the informative words from the summary of the document. The results are evaluated using precision, recall and F1 measure.

***Keywords*— Text summarization, statistical model, F1 measure, key phrase**

I. INTRODUCTION

A gist is a very short summary, ranging in length from a single phrase to a sentence, that captures the essence of a piece of text in much the same way as a title or section heading in a document helps to convey the text's central message to a reader. In this chapter, we present our news story gist system which uses a statistical approach and extractive summarization approach to combine statistical and positional information in order to generate very short news story summaries.

In this paper we proposed a mixed model for gist or headline-styled summary generation. The model is executed in two stages. Initially, the summary of the text document is generated by combining three surface features such as Term frequency of a sentence in the document (TF), Sentence Location in the document (SL) and Centrality of the sentence in the document (CE). In the second stage the informative words are selected from the summary of the document instead from the original document. The informative words are selected from the summary using the statistical model which is a combination of the sentence selection model, content word selection model and text model. From the informative words the key phrases are identified from the original document using the clustering technique. To increase the grammaticality of the sentences in the gist post processing has performed. The rest of the paper is organized as follows: The previous work done in the area short summary generation, headline generation and text summarization are explained in Section 2. Section 3 describes the proposed mixed model for gist generation. Section 4 is dealt with data collection as well as the experimentations. Section 5 is about results analysis, and the conclusions and further research are given in Section 6 and the sample document, the machine generated summary is presented in the chapter 7.

II. LITERATURE REVIEW

Several previous systems were developed to address the need for headline-style summaries. A lossy summarizer that translates news stories into target summaries using the 'IBM- style' statistical machine translation (MT) model was shown in [1]. Conditional probabilities for a limited vocabulary and bigram transition probabilities as gist syntax approximation were incorporated into the translation model. It was shown to have worked surprisingly well with a stand-alone evaluation of quantitative analysis on content coverage. The use of a noisy-channel model and a Viterbi search was shown in another MT-inspired headline summarization system [2]. The method was automatically evaluated by Bi-Lingual Evaluation Understudy [3]. A non statistical system, coupled with linguistically motivated heuristics, using a parse-and- trim approach based on parse trees was reported in [4]. Most of the research has done in extractive summarization methods [5,6,7]. Initially, text summarization process has been studied based on frequent words represent in [8]. First paragraph or first sentences of each paragraph contain topic information proposed in [9]. Query-based summarization is studied in[10]. Maximal Marginal Relevance technique is presented in[11] which is followed in our paper for Telugu text single document summarization. Two-step sentence-extraction method for single-document summarization and multi-document summarization is proposed in [12]. TS using Lexical Chain and WordNet proposed in [13]. The nuclei of the discourse structure tree for a text determine salience of information as in [14].

Text summarization (TS) is a technique which extracts the important information from a text document(s) and produces for particular user or task [15]. In TS, sentences are ranked according to their relevance to the document and extracts the sentences which are more relevant for the document to form a summary until the all the topics in the document are covered without redundancy[16, 17]. The score or the relevance of the sentence can be calculated based different features like syntactic, semantic and structure etc. or combination of these features[18,19].

III. GIST GENERATION PROBLEM

In this paper we viewed the problem of gist generation as extractive summarization problem. The gist has generated mainly using two stages namely text summarization and informative words identification. The different modules in the proposed mixed model are article tokenization, stop word removal, stemming, summary generation, informative word selection using statistical model and post processing for maintaining grammaticality. These modules are explained below.

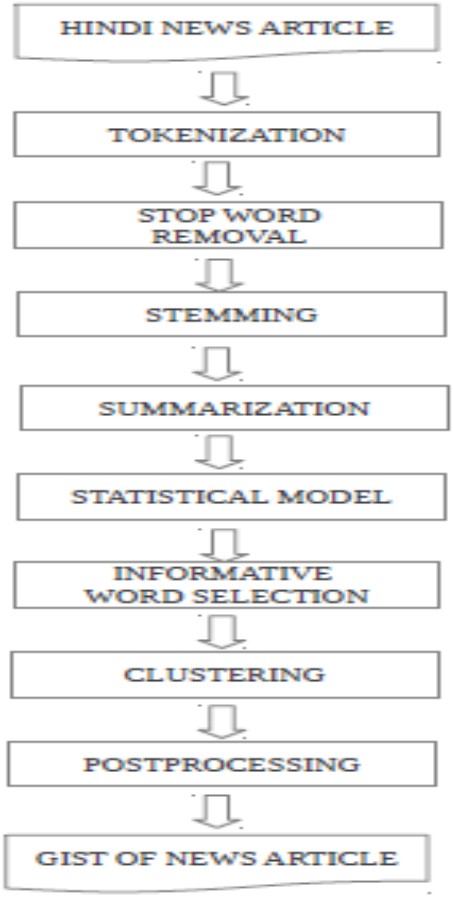

Figure 5.1: The proposed mixed model for gist generation

A. *Pre-processing*

Document need to be pre-processed before processing through the machine. The pre-processing contains removing the unnecessary content from the document which is not useful for TC like punctuation marks, numbers, dates and symbols etc. Secondly, features which can create noise to the summarization process called stop words which are used to give meaning to the sentence need to eliminate. As Hindi is complex morphological variant language, reducing the features of document into their root form can greatly reduces the dimensionality space of the document. Hence features of the document are converted into their root form. After pre-processing the feature space of the document contains only stemmed form of the features.

B. *Text summarization*

Summarization can be broadly divided into two categories namely extractive summarization and abstractive summarization. An extractive summarization is a process of selecting set of sentences from the original document which gives the gist of the document, while an abstractive summarization is a reformulation of the original document probably with new sentences. In this chapter, we have chosen extractive summarization because of simplicity, robustness and domain independence. We have chosen combination of three surface features for summary generation such as Term frequency of a sentence in the document (TF), Sentence Location in the document (SL) and Centrality of the sentence in the document (CE). The formulas for calculating these features and finding the suitable combination of these features are presented below:

i. Term Frequency (TF)

The term frequency of a word in a document is the number of occurrences of the word in that document. This count is normalized to overcome a bias towards longer documents. Term frequency of a word is calculated as follows:

$$tf_i = \frac{c_i}{\sum_k c_k}$$

where 'Ci' is the number of occurrences of a word in the document, and the denominator is the total number of words in the document. The term frequency score of a sentence 'S' is calculated as:

$$Score_{f_1}(S)=\frac{\sum_{k=1}^{n} tf_k}{|S|}$$

where numerator gives the sum of all the words term frequencies in the sentence S and denominator represents the number of words in the sentence.

ii. Sentence Location (SL)

Sentences at the beginning of the texts of news documents give the general information of the document which are suitable to form a gist. The remaining sentences of document are the details about the news which has less importance to include in the gist. Therefore important sentences, which should be included in the gist, are usually located at some particular positions. In order to formalize the sentence location, each sentence is been given a location value Li (Li is equal to i). Then to give higher score to the first sentences, we use the formula mentioned below

$$Score_{f_2}(S)=\frac{R-L_i}{R}$$

which gives the position score of a sentence S.
where 'R' is the number of sentences in the document

iii. Centrality (CE)

The centrality of a sentence implies its similarity to other sentences, which can be measured as the degree of overlapping between sentences to other sentences. If a sentence has high centrality, this sentence introduces many topics of the document. Therefore, high centrality sentences are more preferable in summary than low centrality sentences. To formula to find score of centrality of a sentence 'S' is:

$$Score_{f_3}(S)=\frac{|words\ of\ S \cap words\ of\ remaining\ sentences|}{|words\ of\ S \cup words\ of\ remaining\ sentences|}$$

iv. Summary generation

For a sentence 'S', the weighted score function combine all the feature scores of the sentence as follows.

$$Score(S)=\sum W_i \times Score f_i(S)$$

where Wi represents the weight assigned to feature 'i' to generate the summary that best expresses the gist of the document. For our training dataset, all possible weight combinations between 0 and 1 with an interval of 0.1 between features weights are evaluated. From the empirical evaluations it is concluded that the best weights for TF, CE and SL are 0.2, 0.3 and 0.5 respectively to generate more appropriate summary of the document. Then sentences are ranked according to the above weights assigned to each feature.

IV. MODELS

The informative words are selected from the summary of the document. The informative word selection is based on the statistical model proposed in the previous research work. The statistical model is combinations of three position models namely sentence position model, content word selection model and text model. These models are explained below.

A. 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, what is the likelihood that it would contain the first appearance of a informative word in the key phrase:

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

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

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(Gk|Wj) is a binary feature. This is computed for all sentence positions from 1 to P. Resulting P(G|Posi) represents each sentence position containing one or more informative words.

B. Informative Word Position Model

For each content word Wg, it would most likely first appear at sentence position Posi:

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

In the informative word position model, information was collected for each content word Wg.

C. Text Model

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

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

docTtf(w,j) denotes 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 the text body. For each instance of Gw and Twpair, Gw=Tw.

The following combination of sentence position and text model was used:

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

V. EMPIRICAL EVALUATIONS

A. 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.

B. Evaluation Methods

In this chapter, 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 before summary generation and after summary generation.

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 (Gmachine) and human assigned content words (Ghuman) 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 Gmachine and Ghuman 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. Meanwhile 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.

VI. RESULTS AND DISCUSSIONS

TABLE 5.1: F1 Measures For Eight Possible Combinations

F1 measure	Before summary generation	After summary generation
Sentence position model	0.514	0.573
Informative word selection model	0.352	0.451
Text model	0.486	0.527
Combination of three models	0.602	0.729

The data corpus is evaluated using F1measure for eight possible combinations. Informative words are selected from the original article after preprocessing 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 1.

After the generation of the summary for the given document informative words are selected 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. These results are also specified in the Table 1. The results show the influence of the proposed model on gist generation.

VII. CONCLUSIONS AND FUTURE SCOPE

In this paper, we proposed a mixed model for gist generation for a given Hindi article. The selection of informative words from the summary of the article is more appropriate when compared with the selection of informative words from the whole pre-processed document. The selection of the informative words is 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.

There is a scope for further enhancement of the existing model for increasing the accuracy of the gist using word net and ontology to combine the related words into a single word. The other possibility is by using natural language rules we can improve the informative word selection process.

VIII. SAMPLE ARTICLE

A. **Human generated gist**

चुकंदर से हाई ब्लड प्रेशर के लोगों को फायदा
नाइट्रेट हृदय के लिए बेहतर

B. **Original article**

शोधकर्ताओं का कहना है कि चुकंदर 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 चुकंदर जूस पीया जिससे उनका रक्तचाप 10 एमएमएचजी कम पाया गया| इसका ज़्यादातर असर तीन से छह घंटे तक रहता है, लेकिन अगले दिन भी इसका प्रभाव देखा गया| वैज्ञानिकों का कहना है कि चुकंदर 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| बाट्र्स एंड

द लंदन स्कूल ऑफ़ मेडिसन एंड डेंसिट्री के शोधकर्ता कई वर्षों से ब्लड प्रेशर को कम करने के सिलसिले में चुक़ंदर के प्रभावों पर अध्ययन कर रहे हैं; उनका कहना है कि अभी इस बारे में और काम किए जाने की ज़रूरत है| वो चुक़ंदर 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 चुक़ंदर 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|

C. **Article after summarization**

शोधकर्ताओं का कहना है कि चुक़ंदर 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 चुक़ंदर जूस पीया जिससे उनका रक्तचाप 10 एमएमएचजी कम पाया गया| वैज्ञानिकों का कहना है कि चुक़ंदर 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| 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| 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 चुक़ंदर बहुत फ़ायदेमंद हो सकती हैं"|

D. **Machine generated gist before post-processing**

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

लेगा, उसका हृदय बेहतर तरीक़े से काम करता रहेगा

E. **Machine generated gist after post-processing**

चुक़ंदर के जूस का एक कप पीना हाई ब्लड प्रेशर के शिकार लोगों के लिए

लेगा, उसका हृदय बेहतर तरीक़े से काम करता रहेगा

REFERENCES

[1] Michele Banko, Vibhu Mittal, and Michael Witbrock. 2000. Headline generation based on statistical translation. In ACL-2000, pp. 318-325

[2] David Zajic, Bonnie Dorr, and Richard Schwartz. 2002. Automatic headline generation for news paper stories. In Proceedings of the ACL-2002 Workshop on Text Summarization.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jin Zhu. 2001. IBM research report Bleu: a method for automatic evaluation of machine translation. In IBM Research Division Technical Report, RC22176 (W0109-22).

[4] Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: a parse-and-trim approach to headline generation. In Proceedings of Workshop on Automatic Summarization, 2003.

[5] Kim, J., Kim, J., Hwang, D., 2001. Korean text summarization using an Aggregation Similarity. In: Proc. 5Th Internat. Workshop Information Retrieval with Asian Languages, pp. 111–118.

[6] Nomoto, T., Matsumoto, Y., 2001. A new approach to unsupervised text summarization. In: Proc. ACM SIGIR'01, pp. 26–34.

[7] Wang, J.C., Wu, G.S., Zhou, Y.Y., Zhang, F.Y., 2003. Research on automatic summarization of web document guided by discourse. J.Comput. Res. Develop. 40 (3), 398–405.

[8] Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J. Res. Develop., 159–165.

[9] Edmundson, H.P., 1968. New methods in automatic extraction. J. ACM 16 (2), 264–285.

[10] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999. Summa rizing text documents: Sentence selection and evaluation metrics.In:Proc. ACM-SIGIR'99, pp. 121–128.

[11] Carbonell, J., Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. 21th ACM SIGIR Internat. Conf. on Research and Develop ment in Information Retrieval.

[12] Jung, W., Ko, Y., Seo, J., 2005. In: Automatic Text Summarization Using Two-step Sentence Extraction, LNCS, vol. 3411, pp. 71–81.

[13] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., 1990. Introduction to WordNet: An on-line lexical database (special issue). Internat. J. Lexicogr. 3 (4), 234–245.

[14] Marcu, D., 1996. Building up rhetorical structure trees. In: Proc. 13Th National Conf. on Artificial Intelligence, vol. 2, pp. 1069–1074.

[15] Y. Guo and G. Stylios, "An intelligent summarization system based on cognitive psychology," Information Sciences, Vol. 174, 2005, pp. 1-36.

[16] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 10-17.

[17] D. Marcu, "Discourse trees are good indicators of importance in text," Advances in Automatic Text Summarization, 1999, pp. 123-136.

[18] H. P. Edmundson, "New methods in automatic extracting," Journal of the Association for Computing Machinery Vol. 16, 1969, pp. 264-285

[19] H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of Research and Development, Vol. 2, 1958, pp. 159-165.

