ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 2, February 2017

Design of Automatic Text Summarization Approach for Hindi Text Document Using Semantic Graph and Particle Swarm Optimization

[1] Vipul Dalal [2] Manisha Waze [3] Dr. Latesh Malik
[1] Research Scholar, CSE Department
G.H.Raisoni College of Engineering, Nagpur.
[2] Associate Professor, Computer Department,
Govt. College of Engineering, Nagpur

***Abstract: --* Automatic text summarization is the process of summarizing given document using intelligent algorithms. Many techniques have been suggested by researchers in past for summarization of English text. Not much work is found in the literature for summarization of Hindi text even though Hindi is an official language of India. In this paper, we propose a design for summarizing Hindi text based on semantic graph of the document using Particle Swarm Optimization (PSO) algorithms. The subject-object-verb (SOV) triples are extracted from the document. These triples are used to construct semantic graph of the document. A trained classifier using PSO algorithm generates semantic sub-graph which is then used to obtain document summary. The approach is under implementation phase and expected to give better results as compared to traditional summarizers.**

***Keywords*— text mining, text summarization, feature extraction.**

I. INTRODUCTION

Hindi is official language of India. It is native language of more than 258 million people in India. The use of Hindi documents in various fields is increasing rapidly. Text summarization can help readers by providing a glimpse of the document. To automate the process of summarization, researchers generally rely on a two phase process. First, key textual elements, e.g., keywords, clauses, sentences, or paragraphs are extracted from text using linguistic and statistical analyses. In the second step, the extracted text may be used as a summary. Such summaries are referred to as "extracts". Another approach called "abstractive summarization" consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. Such abstracts may or may not contain the sentences from the original document. In the survey of literature we found very little documented work for summarizing Hindi text [1]. So, in this paper we propose a design of semantic graph based approach for summarizing Hindi text using PSO algorithm.

In either case, preprocessing of input text plays vital role. The preprocessing phase generally consists of three sub-steps: (i) parsing the input text, (ii) identifying and labeling individual tokens as noun, proper noun, verb, adjective, etc. (iii) stop word removal. It is normally followed by feature extraction phase.

The rest of this paper is organized as follows. Section II explains related work done for automatic text summarization. Section III gives detailed explanation of the proposed approach. Section IV discusses the results obtained by the proposed approach. Section V gives conclusion.

II. RELATED WORK

A lot of efforts have been taken by researchers to extract a rich set of features required for summarization process. Joel Larocca Neto et al [14] reviewed and employed a large variety of features. They used mean TF-ISF, sentence length, sentence position, similarity to title, similarity to keywords, sentence-to-sentence cohesion, sentence-to-centroid cohesion, occurrence of proper noun, and occurrence of anaphor.

They used Naïve bayes and C4.5 machine learning algorithms to train a classifier using these features.
Aysun Güran et al [15] analyzed the performances of a feature-based and two semantic-based text summarization algorithms on Turkish corpus. They used title, cue words, paragraph location, proper nouns, term frequency, adverbs, numeric literals, average length, keywords, and quotations as their feature vector. Alkesh Patel et al [16] described an algorithm for language independent generic extractive summarization for single document. The algorithm is based on structural and statistical (rather than semantic) factors. They used in their algorithm noun feature vector, document feature vector, theme feature vector, and location feature.

Nguyen Quang Uy et al [17] presented an application of Genetic Programming to the problem of Automatic Text Summarization. Genetic Programming was used to evolve the function that ranks the sentences in a document based on their importance. The summary was extracted by selecting the sentences that have the highest rankings. They used location of paragraph, location of sentence, length of sentence and content word frequency as their feature vector.

Massih R. Amini et al [18] presented an approach for Single Document Summarization based on a Machine Learning ranking algorithm. They employed cue words, frequency, title keywords, location and length of sentence feature vector to train a classifier.

Khosrow Kaikhah [19] presented a novel technique for summarizing news articles using neural networks. A neural network was trained to learn the relevant features of sentences that should be included in the summary of the article. He used features like paragraph follows title, paragraph location, sentence location, first sentence in paragraph, sentence length, number of thematic words in sentence, and number of title words in sentence to train a neural network.

Kamal Sarkar [20] presented a method for Bengali text summarization which extracts important sentences from a Bengali document to produce a summary. He used TF-IDF, sentence position, and sentence length as the feature space. The summary was generated by simply selecting the top scored sentences from the original text.

Albaraa Abuobieda M. Ali et al [21] presented a feature selection method using (pseudo) Genetic probabilistic-based Summarization (PGPSum) model for extractive single document summarization. The proposed method, working as features selection mechanism, was used to extract the weights of features from texts. Then, the weights were used to tune features' scores in order to optimize the summarization process. In this way, important sentences were selected for representing the document summary. They used five simple features: title feature, sentence length, sentence position, numerical data and thematic word.

The extractive automatic text summarization work involving bio-inspired algorithms such as PSO is as follows. M. S. Binwahlan et al [2] introduced a work for feature selection. They exploited five features regarding to text summarization and the PSO was used to train the system to obtain the weights of each feature. These weights have been employed in their next work [3] to generate the best summary. The results shown that, the proposed PSO method generate summaries which are 43% similar to the manually generated summaries, while MS-Word summaries are 37% similar.

Albaraa Abuobieda M. Ali et al [21] presented a feature selection method using (pseudo) Genetic probabilistic-based Summarization (PGPSum) model for extractive single document summarization. The proposed method, working as features selection mechanism, was used to extract the weights of features from texts. Then, the weights were used to tune features' scores in order to optimize the summarization process. In this way, important sentences were selected for representing the document summary. The results showed that, their PGPSum model outperformed Ms-Word benchmarks by obtaining a similarity ratio closest to human benchmark summary.

III SUMMARIZATION OF INDIAN TEXT

An algorithm for language independent generic extractive summarization for single document is proposed by Patel et al [5]. The algorithm is based on structural and statistical parameters. The proposed algorithm was performed over a single-document summarization for English, Hindi, Gujarati and Urdu documents. Naresh Kumar Nagwani et al [6] designed and implemented a frequent term based text summarization algorithm. The designed algorithm works in three steps. In the first step the document which is required to be summarized is processed by eliminating the stop word and by applying the stemmers. In the second step term-frequent data is calculated from the document and frequent terms are selected, for these selected words the semantic equivalent terms are also generated. Finally in the third step all the sentences in the document, which are containing the frequent and semantic equivalent terms, are filtered for summarization.

Kamal Sarkar [7] proposed an approach which extracts important sentences from a Bengali document to produce a summary. The sentences were ranked based on two important features: thematic term and position.
Upendra Mishra et al [8] developed a new stemmer named as “Maulik” for Hindi Language. The stemmer can be used in the preprocessing phase of summary generation. Vishal Gupta et al [9] proposed preprocessing phase for Punjabi text summarization. They mainly apply stop word removal, noun stemming and cue phrase detection.

IV PROPOSED APPROACH

Jurij Leskovec et al [10] presented a method for summarizing document by creating a semantic graph of the original document and identifying the substructure of such a graph that can be used to extract sentences for a document summary. Our proposed design is based on this concept, but is applied on Hindi text. Instead of training SVM classifier, we are using Particle Swarm Optimization (PSO) to train the classifier. The PSO approach is well known for its optimization capabilities.

The pre-processing phase and construction of feature vector can be summarized as follows:

Step 1- Parse the document using a parser for POS tagging and dependency tagging.
Step 2- Identify subjects, objects and verbs from each sentence.
Step 3- Extract subject-object-verb (SOV) triples from each sentence.
Step 4- Extract simple document level features like sentence length, sentence similarity, word position, TF-ISF, word frequency, etc.
Step 5- Extract semantic level features like POS tag, dependency tag, SOV tag, etc.
Step 6- Construct document’s semantic graph.
Step 7- Extract graph level features like pagerank, authority, hub, number of incoming links, number of immediate neighbors, etc.
Step 8- Combine all features to construct n-dimensional feature vector.
Step 9- Normalize the feature vector.
Step 10 – Repeat these steps for all the documents in the training set.

This normalized feature vectors in the training set can be used for training a classifier. It could be Bayes classifier, Neural network, SVM or a bio-inspired classifier. For this work, we have employed the parser developed by Siva Reddy [22]. The parser’s accuracy as they specify is about 78%. Figure 1 shows flow of preprocessing phase and construction of feature vector.

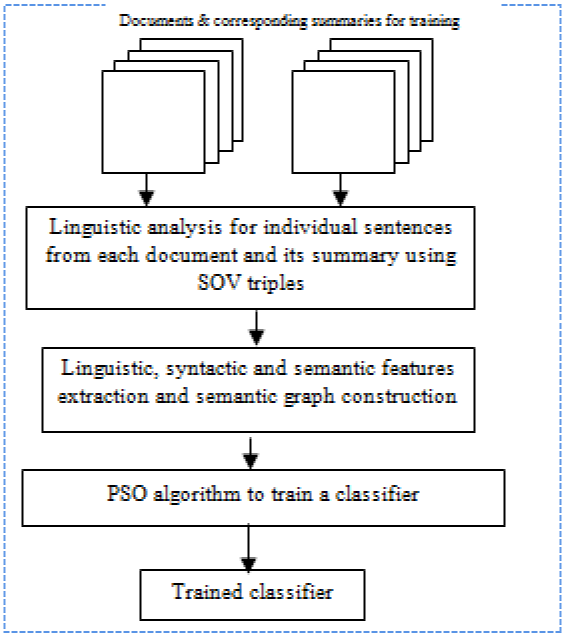
Documents & corresponding summaries for training
Linguistic analysis for individual sentences from each document and its summary using SOV triples
Linguistic, syntactic and semantic features extraction and semantic graph construction
PSO algorithm to train a classifier
Trained classifier

Figure 1. Off-line training phase of the summarizer

Once the classifier is trained, summary extraction can be done as follows:
Step 1 - Use the trained classifier to derive sub-graph structure from the semantic graph of the input document.

Step 2- Generate summary using the sub-graph obtained from the classifier.
A semantic graph of a sample document can be visualized as shown in figure 3.

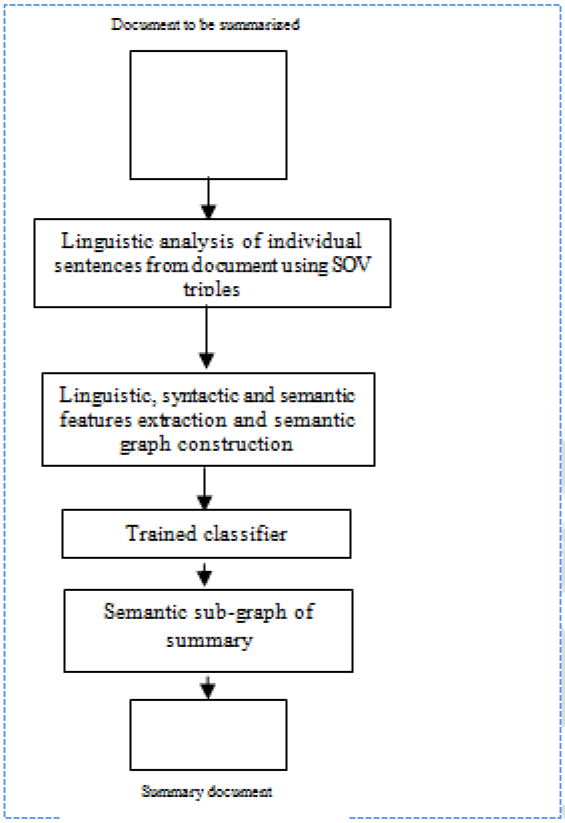
Document to be summarized
Linguistic analysis of individual sentences from document using SOV triples
Linguistic, syntactic and semantic features extraction and semantic graph construction
Trained classifier
Semantic sub-graph of summary
Summary document

Figure 2. Real time summarization phase

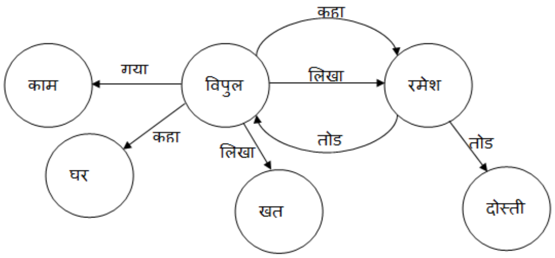

Figure 3. A semantic graph of a sample document.

V CONCLUSION

In this paper we presented a bio-inspired text summarization approach based on semantic graph of input document for Hindi text. The traditional summarizer relies upon sentence score obtained using various features but does not optimally select the summary sentences. Our proposed approach uses PSO to select the summary sentences optimally. The approach is under implementation phase and expected to give better results as compared to traditional summarizers.

REFERENCES

[1] Vipul Dalal, Dr. Latesh Malik, “A Survey of Extractive & Abstractive Text Summarization”, 6th International Conference on Emerging Trends in Engineering & Tecnology (ICETET), 2013

[2] M. S. Binwahlan, Salim, N., & Suanmali, L., "Swarm based features selection for text summarization," International Journal of Computer Science and Network Security IJCSNS, vol. 9, pp. 175-179, 2009b.

[3] M. S. Binwahlan, et al., "Swarm Based Text Summarization," in Computer Science and Information Technology – Spring Conference, 2009. IACSITSC '09. International Association of, 2009, pp. 145-150.

[4] Albaraa Abuobieda M. Ali, Naomie Salim, Rihab Eltayeb Ahmed, Mohammed Salem Binwahlan, Ladda Sunamali, Ahmed Hamza, “Pseudo Genetic And Probabilistic-Based Feature Selection Method For Extractive Single Document Summarization”, Journal of Theoretical and Applied Information Technology, 15th October 2011. Vol. 32 No.1, ISSN: 1992-8645, E-ISSN: 1817-3195.

[5] Alkesh Patel, Tanveer Siddiqui, U. S. Tiwary, “A language independent approach to multilingual text summarization”, Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007 - Copyright C.I.D. Paris, France

[6] Naresh Kumar Nagwani, Shrish Verma, “A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm”, International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011.

[7] Kamal Sarkar, “Bengali Text Summarization By Sentence Extraction”

[8] Upendra Mishra, Chandra Prakash, “MAULIK: An Effective Stemmer for Hindi Language” International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397, Vol. 4 No. 05 May 2012

[9] Vishal Gupta, Gurpreet Singh Lehal, “Preprocessing Phase of Punjabi language Text Summarization”

[10] Jurij Leskovec, Natasa Milic-Frayling, Marko Grobelnik, “Extracting Summary Sentences Based on the Document Semantic Graph” Microsoft Research, Microsoft Corporation

[11] Regina Barzilay, Michael Elhadad. “Using Lexical Chains for Text Summarization”. In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97). Madrid: ACL, 1997. 10-17.

[12] Kavita Ganesan, ChengXiang Zhai, Jiawei Han, “Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions”.

[13] Eduard Hovy and Chin-Yew Lin. “Automated Text Summarization in SUMMARIST”. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. MIT Press.

[14] Joel Larocca Neto, Alex A. Freitas Celso, A. A. Kaestner, “Automatic Text Summarization using a Machine Learning Approach”, Advances in Artificial Intelligence, Lecture Notes in Computer Science Volume 2507, 2002, pp 205-215

[15] Aysun Güran, Eren Bekar, Selim Akyokuş, “A Comparison of Feature and Semantic-Based Summarization Algorithms for Turkish”, INISTA 2010, International Symposium on Innovations in Intelligent Systems and Applicaitons, 21-24June 2010, Kayseri & Cappadocia,TURKEY

[16] Alkesh Patel, Tanveer Siddiqui, U. S. Tiwary, "A language independent approach to multilingual text summarization", Conference RIAO2007, Pittsburgh PA, U.S.A., (2007).

[17] Nguyen Quang Uy, Pham Tuan Anh, Truong Cong Doan, Nguyen Xuan Hoai, “A Study on the Use of Genetic Programming for Automatic Text Summarization”, 2012 Fourth International Conference on Knowledge and Systems Engineering

[18] Massih R. Amini, Nicolas Usunier, and Patrick Gallinari, “Automatic Text Summarization Based onWord-Clusters and Ranking Algorithms”, ECIR 2005, LNCS 3408, pp. 142–156, 2005.Springer-Verlag Berlin Heidelberg 2005

[19] Khosrow Kaikhah, “Automatic Text Summarization with Neural Networks”, SECOND IEEE International Conference On Intelligent Systems, June 2004.

[20] Kamal Sarkar, “Bengali Text Summarization By Sentence Extraction“.

[21] Albaraa Abuobieda M. Ali, Naomie Salim, Rihab Eltayeb Ahmed, Mohammed Salem Binwahlan, Ladda Suanmali, Ahmed Hamza, “Pseudo Genetic And Probabilistic-Based Feature Selection Method For Extractive Single Document Summarization”, Journal of Theoretical and Applied Information Technology 15th October 2011. Vol. 32 No.1.

[22] Reddy Siva, Natural Language Processing Tools. December. 2012 URL: http://sivareddy.in/downloads.