Data Clustering Approach for Automatic Text Summarization of Hindi Documents using Particle Swarm Optimization and Semantic Graph

Vipul Dalal, Latesh Malik

Abstract: ***Automatic text summarization is a process of describing important information from given document using intelligent algorithms. A lot of methods have been proposed by researchers for summarization of English text. Automatic summarization of Indian text has received a very little attention so far. In this paper, we have proposed a data clustering approach for summarizing Hindi text using semantic graph of the document and Particle Swarm Optimization (PSO) algorithm. PSO is one of the most powerful bio-inspired algorithms used to obtain optimal solution. The subject-object-verb (SOV) triples are extracted from the document. These triples are used to construct semantic graph of the document and finally clustered into summary and non-summary groups. A classifier is trained using PSO algorithm which is then used to obtain document summary.***

Keywords: ***bio-inspired algorithms, text mining, text summarization, semantic graph, PSO, data clustering***

I. INTRODUCTION

Hindi is national 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 allows readers to get a gist of a given document. The process of automatic text summarization consists of two phases. In the first phase, called “pre-processing phase”, key textual elements, such as keywords and clauses are extracted from the given text. This requires linguistic and statistical analysis of the text. In the second phase, the extracted text is used as a summary. Such summaries are called “extracts” and this type of technique is called “extractive summarization”. Another approach is called “abstractive summarization”. In this approach the original text is interpreted and described in fewer sentences. Here linguistic methods are used to examine and interpret the text. The new concepts and expressions are found which can describe the text in a new shorter form such that it conveys the most relevant information from the original text. Such abstracts may or may not contain the sentences from the original document. Extractive summarization is shallow approach and is easy to implement whereas abstractive summarization needs deep understanding and analysis of the document and involves some elements of Natural Language Generation (NLG), so it is more complex to implement. Our proposed approach extracts summary sentences from the input document only but analyzes semantic relationships of the document elements.

In the survey of literature we found very little documented work for summarizing Hindi text [1]. So, in this paper we have proposed a semantic graph based approach for summarizing Hindi text using PSO algorithm. The rest of the paper is organized as follows. In section 2, related work based on bio-inspired techniques is explained. Section 3 explains related work for Indian languages especially for Hindi. Section 4 explains our proposed approach for Hindi document. Experiment and results are discussed in section 5. Finally, section 6 concludes the proposed approach.

II. SUMMARIZATION USING BIO-INSPIRED METHODS

The extractive automatic text summarization work based on bio-inspired algorithms is as follows.

M. S. Binwahlan et al [2] introduced an approach for feature selection. In their approach five features related to text summarization were used and the PSO was employed to make the system learn to obtain the weights of each feature. These weights are used in their next work [3] to generate the summary. The authors claimed that, their PSO method can generate summaries that are 43% similar to the human generated summaries, whereas summaries generated by MS-WORD are 37% similar.

Albaraa Abuobieda M. Ali et al [4] proposed a feature selection approach based on (pseudo) Genetic probabilistic-based Summarization (PGP Sum) model. This model was used for generating extractive summary of single document. Their method was employed as features selection mechanism and was used to obtain the weights of features from texts. These weights were used to obtain tuned scores for the features and to optimize the summarization process. The document summary was represented using these important sentences. The authors claimed that, their PGP Sum model is better than Ms-Word benchmarks as the similarity ratio is close to human benchmark summary.

III. SUMMARIZATION OF INDIAN TEXT

An approach for generic extractive summarization for single document was proposed by Patel et al [5]. Various structural and statistical parameters were used in their method. The algorithm was claimed to be language independent and it was applied to generate single-document summary for English, Hindi, Gujarati and Urdu documents. Naresh Kumar Nagwani et al [6] developed a frequent term based text summarization algorithm. There are three steps in the algorithm. The first step processes the input document,

Revised Version Manuscript Received on May 22, 2017.

Vipul Dalal, Research Scholar, Department of Computer Science & Engineering, G.H. Raisoni College of Engineering, Nagpur (Maharashtra)-440016, India, E-mail: vipul.dalal@vit.edu.in

Dr. Latesh Malik, Department of Computer Science & Engineering, Government College of Engineering, Nagpur (Maharashtra) - 440016, India, E-mail: latesh.gagan@gmail.com

Eliminates the stop words and applies the stemmers to obtain root words. In the second step frequent terms are obtained from the document. These frequent terms are filtered to get the top frequent terms that are further considered. For these terms the semantic equivalent terms are also extracted. A last the third step filters all the sentences in the input document, that contain the frequent and semantic equivalent terms, and generates summary.

Kamal Sarkar [7] proposed an extractive approach for Bengali text summarization. A ranking was generated for the sentences based on thematic term and position features. Upendra Mishra et al [8] designed a stemmer named "Maulik" for Hindi Language. This stemmer may be used to obtain root words in the preprocessing phase of summarization. Vishal Gupta et al [9] suggested preprocessing phase for Punjabi text summarization. In this work, they applied stop word removal, noun stemming and cue phrase detection.

IV. PROPOSED APPROACH

We found from the literature survey that very little work is done for summarization of Hindi text. In this paper, we have proposed an approach based on [10]. 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. Our approach can be outlined as follows:

1) Preprocess a set of training documents as well as the corresponding summaries to extract SOV triples from each sentence.
2) Construct semantic graphs for the training documents and their corresponding summaries using the extracted SOV triples.
3) Train PSO classifier to learn semantic sub-graph structure of the summaries from the semantic graph of the corresponding training documents. This procedure is depicted in Figure 1

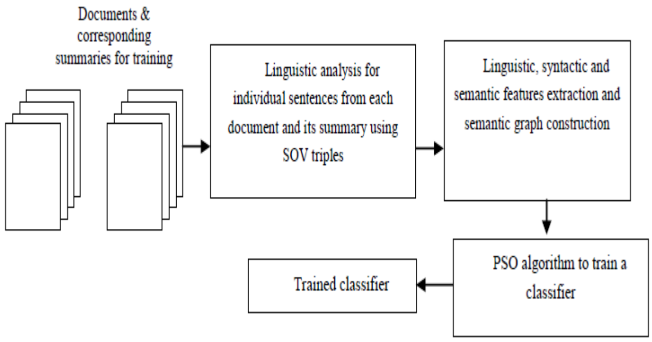
Documents & corresponding summaries for training
Linguistic analysis for individual sentences from each document and its summary using SOV triples
Linguistic, syntactic and semantic features extraction and semantic graph construction
PSO algorithm to train a classifier
Trained classifier

Figure 1. Offline Training Phase

4) Preprocess the input document to extract SOV triples and to construct its semantic graph.
5) Use the trained classifier to derive sub-graph structure from the semantic graph of the input document.
6) Generate summary using the sub-graph obtained from the classifier.

This procedure is depicted in Figure 2.

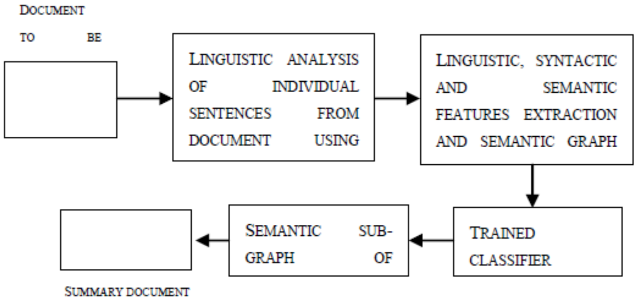
DOCUMENT TO BE
LINGUISTIC ANALYSIS OF INDIVIDUAL SENTENCES FROM DOCUMENT USING
LINGUISTIC, SYNTACTIC AND SEMANTIC FEATURES EXTRACTION AND SEMANTIC GRAPH
TRAINED CLASSIFIER
SEMANTIC SUB-GRAPH OF
SUMMARY DOCUMENT

Figure 2. Real Time Summarization Phase

V. EXPERIMENT AND RESULTS

The proposed approach is implemented using Java platform. A set of 80 Hindi documents along with their summaries were selected for training purpose. In the preprocessing phase, total 1550 SOV triples were extracted to form the training set. A feature vector comprising of total 144 features was obtained for each SOV triple of each training document. The selected features can be categorized as follows:

Linguistic features – This includes POS tags, dependency tags, subject-object-verb tags, word depth in the dependency tree, etc.

Semantic Graph features – This includes page rank, hub, authority, number of incoming links, number of out going links, number of direct neighbors, number of indirect neighbors, etc.

Document Discourse Structure features – This includes sentence length, word position, word frequency, tf-isf, sentence similarity, etc.

Mainly the graph based features allow our proposed approach to perform deep semantic analysis of document elements. So, our approach gives a good compromise between the simplicity of extractive summarization and the human-like summary generation capability of abstractive summarization.

For the sack of simplicity, co-reference resolution and anaphora resolution were ignored and done manually. After forming the feature set, the PSO algorithm was run to obtain the centroids. The swarm was empirically considered converged if there is no improvement for 10 consecutive iterations or if "swarm size X dimensions" (i.e. 126X144, in this case) number of iterations is executed. The final global best position gives near optimal centroids. The feature vector of each SOV triple in the input document is then compared with these centroids and appropriate label is assigned to each triple. The sentences with at least one SOV triple labeled as part of reduced graph, are then included in the final summary of the document.

The unavailability of benchmark for Hindi summarization makes evaluation of our approach difficult. Therefore, the extracted summary was compared with human extracted summary. The system's performance was measured using precision, recall, F1 score and G score.

$$precision=\frac{no\ of\ \ summary\ \ sentences\ extracted\ \ that\ match\ \ with\ \ human\ exctracted\ \ summary}{total\ number\ of\ \ sentences\ extracted} \qquad (1)$$

$$recall=\frac{no\ of\ \ summary\ \ sentences\ extracted\ \ that\ match\ \ with\ \ human\ exctracted\ \ summary}{no\ \ of\ \ actual\ summary\ \ sentences\ in\ human\ \ extracted\ \ summary} \qquad (2)$$

$$F1=2\,\frac{precision\cdot\ recall}{precision+recall} \qquad (3)$$

$$G=\sqrt{precision\cdot\ recall} \qquad (4)$$

The performance of the proposed approach is given in table 1. Higher value of recall indicates more sensitivity of the approach as compared to the accuracy or the precision.

Table 1. Performance metrics for the proposed approach.

recall	60
precision	42.86
F1 score	50.01
G score	50.71

VI. CONCLUSION

In this paper we have presented a bio-inspired text summarization approach based on semantic graph of input document for Hindi text. The traditional summarizers rely upon sentence score obtained using various features but do not optimally select the summary sentences. Our proposed approach uses PSO to select the summary sentences optimally. The approach gives reasonably good performance. The adequacy of the approach can be improved if anaphora resolution and co-reference resolution are integrated in the preprocessing phase.

REFERENCES

1. Luhn, H.P., 1958. “The automatic creation of literature abstracts”. IBM J. Res. Develop., 2: 159-165.
2. P. B. Baxendale, "Machine-made index for technical literature: an experiment," IBM J. Res. Dev., vol. 2, pp. 354-361, 1958.
3. Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2):264-285.
4. Lin, C.Y. 1999. “Training a selection function for extraction”. Proceedings of the 18th Annual International ACM Conference on Information and Knowledge Management, pp:55-62.
5. Massih R. Amini, Nicolas Usunier, and Patrick Gallinari, "Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms", ECIR 2005, LNCS 3408, pp. 142–156, (2005).
6. Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords", International Arab Journal of e-Technology, Vol. 1, No. 4, June, pp. 164-168, (2010).
7. Jade Goldstein, Jaime Carbonell. “SUMMARIZATION: (1) USING MMR FOR DIVERSITY- BASED RERANKING AND (2) EVALUATING SUMMARIES”. Carnegie Group Inc.'s Tipster III Summarization Project
8. Aysun Güran, Eren Bekar, Selim Akyokuş “A Comparison of Feature and Semantic-Based Summarization Algorithms or Turkish”. INISTA 2010, International Symposium on Innovations in Intelligent Systems and Applicaitons, 21-24June 2010, Kayseri & Cappadocia,TURKEY.
9. Ono, K., Sumita, K., and Miike, S. (1994). “Abstract generation based on rhetorical structure extraction.” In Proceedings of Coling '94, pages 344{348, Morristown,NJ, USA.
10. Marcu, D. (1998a). “Improving summarization through rhetorical parsing tuning”. In Proceedings of The Sixth Workshop on Very Large Corpora, pages 206-215, pages 206,215, Montreal, Canada.
11. Giuseppe Carenini and Jackie Chi Kit Cheung, “Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality”.
12. Pierre-Etienne Genest, Guy Lapalme, “Framework for Abstractive Summarization using Text-to-Text Generation”, Workshop on Monolingual Text-To-Text Generation, pages 64–73,Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 64–73,Portland, Oregon, 24 June 2011. c 2011 Association for Computational Linguistics.
13. Vipul Dalal, Dr. Latesh Malik.: “A Survey of Extractive & Abstractive Text Summarization”, 6th International Conference on Emerging Trends in Engineering & Tecnology (ICETET), 2013
14. 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.
15. M. S. Binwahlan, Salim, N., & Suanmali, L.: “Swarm Based Text Summarization”, Computer Science and Information Technology – Spring Conference, 2009. IACSITSC '09. International Association of, 2009, pp. 145-150.
16. 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.
17. 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
18. 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.
19. Kamal Sarkar.: “Bengali Text Summarization By Sentence Extraction”
20. 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
21. Vishal Gupta, Gurpreet Singh Lehal.: “Preprocessing Phase of Punjabi language Text Summarization”
22. Jurij Leskovec, Natasa Milic-Frayling, Marko Grobelnik.: “Extracting Summary Sentences Based on the Document Semantic Graph, Microsoft Research, Microsoft Corporation
23. 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.
24. Kavita Ganesan, ChengXiang Zhai, Jiawei Han.: “Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions”.
25. 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.
26. Udo Hahn, Inderjeet Mani. : “The Challenges of Automatic Text Summarization”, IEEE Computer Society Press Los Alamitos, CA, USA, Volume 33 Issue 11, November 2000, Page 29-36 ISSN:0018-9162.
27. Chetana Thaokar, Latesh Malik, “Test Model for Summarizing Hindi Text using Extraction Method”, Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).
28. Reddy Siva. Natural Language Processing Tools. December. 2012 URL: http://sivareddy.in/downloads