ISSN:0975-9646

A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique

M. S. Patil [#1], M. S. Bewoor[*2], S. H. Patil[#3]

[#1]*M.Tech Computer Department, Bharati Vidyapeeth University College of Engineering Pune, India*

[*2]*Associate Professor Computer Department, Bharati Vidyapeeth University College of Engineering Pune ,India*

[#3] *Professor Computer Department, Bharati Vidyapeeth University College of Engineering Pune, India*

***Abstract* — Usually, presence of the same information in multiple documents is the main problem faced in effective information access. Instead of this redundant information thus accessed or retrieved, users are interested in retrieving information that addresses one or other several aspects. In such situation, text summarization proves to be very useful. Not only in Information retrieval, but it is an extremely active research topic in other fields like natural language processing and machine learning. Text summarization is a process of extracting content from a document and generating summary of that document thus presenting important content to user in a relatively condensed form. In this paper, study of several extractive text summarization approaches is made and an effective text summarization method is proposed. This method is based on Support-Vector-Machine (SVM). Proposed system tries to improve the performance and quality of the summary generated by the clustering technique by cascading it with SVM.**

Keywords— **clustering, document summarization, extractive text summarization, machine learning, SVM.**

I. INTRODUCTION

Text summarization is gaining much importance nowadays. One reason for this is, recently due to the enormous growth in information, need for automatic text summarization has increased. Hence, it is usual that users trying to retrieve the documents or information face the problem of responses of hundreds or thousands of retrieved documents

Also, retrieved documents have most of the redundant information. Hence, it is not easy for users to manually summarize the large number of documents. So it is desirable to have a system that could summarize these documents. Text summarization satisfies this user's need by summarizing the text documents.

Another reason is that the obvious overlap of text summarization with information extraction, and connections from summarization to both automated question answering and natural language generation, suggest that summarization is actually a part of a larger picture. These fields offer a huge scope to concise and compact the information enabling the user to decide by mere check at snippets of each link. Hence summarization is an important activity in the analysis of a huge volume of text documents. The purpose of the text summarization is to present the main idea in a document in less space so that user will not have to waste time in reading the whole document. Text summarization is nothing, but a text that is produced from one or more documents that convey the user all the important information in the original text.

The objective and approach of summarization of documents explain the kind of summary that is generated. The summary generated reveals the salient and shared information of its documents. Summary, so generated, must be query dependent. This is so because of huge growth of information on internet has led to the use of IR systems which work on search engines. Search engines retrieve the documents based on query. Hence, it is important that the summarization is also query dependent.

Broadly, text summarization can be classified into two types:

Extractive: Extractive summarization methods simplify the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. This type of summarization picks out the most relevant sentences in the document thereby maintaining the low redundancy in summary.

Abstractive: Abstractive summarization may compose novel sentences, unseen in the original sources.

II. RELATED WORK

This section deals with the literature review of the extractive text summarization techniques.

The paper [3] deals with an automatic trainable summarization procedure which is based on the application of machine learning techniques. This summarizer can be obtained by applying trainable machine learning algorithm for collection of documents as well as their summaries. Here the sentences of each document are modelled as vectors of features extracted from the text. For this, summarization task can be considered as a two-class classification problem, where a sentence is labelled as "correct" if it belongs to the extractive reference summary, or as "incorrect" otherwise. The patterns which lead to the summaries are expected to be learned by the trainable summarizer. This is done by identifying relevant feature values which are most correlated with the classes "correct" or "incorrect". The "learned" patterns are used to classify

each sentence of that document as a "correct" or "incorrect" sentence, producing an extractive summary for a new document given to the system. In this paper, author has proposed a trainable summarizer that uses a large variety of features, some of them employing statistics-oriented procedures and others using linguistics-oriented ones. For the classification task, they have used two different well known classification algorithms, namely the Naive Bayes algorithm and the C4.5 decision tree algorithm. The performance of these procedures was compared with the performance of two non-trainable, baseline methods. In this comparison, the trainable method using Naive Bayes classifier significantly outperformed all the baseline methods.

In [6], author has proposed a Single Document Summarization approach based on Machine Learning ranking algorithm. For this, set of features are used in order to produce a vector of scores for each sentence in a document. To make a global combination of these scores, a classifier is trained. Here features used are built upon word-clusters. These word-clusters are nothing but groups of words that co-occur with each other. Also, they can serve to expand a query and also to enrich the representation of the sentences in the documents. Experiments conducted show that the learning algorithms perform better than the non-learning systems. Using training set of documents and their associated summaries, these classification approaches usually train a classifier in order to distinguish between summary and non-summary sentences.

In this paper, author has proved that the ranking algorithm outperforms the classification algorithm on both the datasets used. Also the difference of performance between the two algorithms depends on the nature of the collection. The hypothesis on the dataset made by a linear ranker is that hyper plane separates relevant and irrelevant sentences of a given document in the feature space.

Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda in [12] has proposed extraction of sentences that contain important information from a document based on Support Vector Machine. According to author it is the technique for text summarization is the key to the automatic generation of summaries that will generate similar summaries written by humans. Integration of Heterogeneous pieces of information must be done to achieve such extraction. One approach is parameter tuning by machine learning. It has attracted a lot of attention. In this paper, Support Vector Machine based method of sentence extraction is proposed. To confirm the method's performance, author has conducted experiments that compare their method to three existing methods decision tree learning, boosting, lead, and SVM. Results on the Text Summarization Challenge (TSC) corpus show that this method offers the highest accuracy. Moreover, it is clarified that the different features effective for extracting different document genres.

In [11] the experimental results performed shows that SVMs consistently achieve good performance on text categorization tasks, outperforming existing methods. SVM has an ability to generalize well in high dimensional feature spaces. This characteristic eliminates their need for feature selection there by making the application of text categorization considerably easier. Robustness is another advantage of SVMs over the conventional methods. Experiments show the SVM's good performance as they avoid catastrophic failure, which is observed with the conventional methods for some tasks. In addition to this, parameter tuning is not necessary for SVMs as they can find good parameter settings automatically. All these properties of SVM makes it very promising as well as easy to use in order to learn text classifiers from examples.

In [5] author has compared the performance of neural networks and Support Vector Machines for text summarization. These both techniques have ability to discover nonlinear data. Also they are effective for large datasets. Results of the experiments conducted by author shows that neural network are slower than SVM in large datasets.

III. Proposed System Architecture

Text summarization consists of main three steps pre-processing step, processing step and summary generation. Pre-processing step obtains a structured representation of the original text. Processing step deals with the algorithm that transforms the text into summary. Summary generation obtains full summary from summary structure obtained from processing step.

Figure 1 shows the proposed system architecture. Proposed system will consists of following phases.

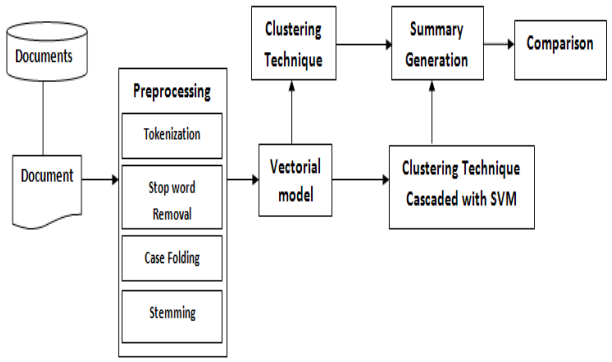

Fig. 1 System Architecture of Proposed System

i. Pre-processing step:

It consists of NLP phases like tokenization, stop word removal, case folding and stemming.

a. Tokenization: In this phase sentences are divided into streams of individual tokens that are differentiated by spaces.

b. Stop Word Removal: Stop words are the words that occur frequently. These words must be eliminated because they influence the sentences that contain these words.

c. Case Folding: It is to reduce all letters to lower case.

d. Stemming: Stemming is the process of reducing derived or inflected words to their stem, base or root form—generally a written word form.

ii. Clustering Technique

This phase will use the clustering algorithm so as to create the summary.

iii. Clustering Technique Cascaded with Support-Vector-Machine.

For this phase, an algorithm will be generated containing clustering technique cascaded with the machine learning technique i.e Support-Vector-Machine (SVM).

iv. Summary Generation

Summary of the text document will be generated using two techniques, namely the clustering technique and clustering technique cascade with Support Vector Machine.

v. Comparison

Following metrics will be used to evaluate the summaries generated:

1) Semantic Gap
2) Misclassification cost
3) Purity
4) Cluster entropy
5) V-measure

After evaluating the metrics for summaries generated, comparison of the summaries will be done.

IV. CONCLUSIONS

Literature review shows that SVMs are the universal learners. In their basic form, linear threshold function is learnt by SVMs. One remarkable property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space. Based on the margin with which they separate the data, SVMs measure the complexity of hypotheses and not the number of features. Hence, if SVM is used with any of the clustering algorithm it will definitely improve the quality of the summary generated by the clustering technique alone.

REFERENCES

[1] H. Jing and K. McKeown. "Cut and paste based text summarization". In Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pages 178--185, 2000

[2] Jiaming Zhan and Han-tong Loh "Using Redundancy Reduction in Summarization to Improve Text Classification by svms" Journal of information science and engineering 25, 591-601 (2009)

[3] Joel Larocca Neto Alex A. Freitas Celso A. A. Kaestner, "Automatic Text Summarization using a Machine Learning Approach" SBIA '02 Proceedings of the 16th Brazilian Symposium on Artificial Intelligence Advances in Artificial Intelligence

[4] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA – International Journal of Computing Science and Communication Technologies, vol. 2, no. 1, Jul. 2009

[5] Keivan Kianmehr, Shang Gao, Jawad Attari, M. Mushfiqur Rahman,Kofi Akomeah, Reda Alhajj, Jon Rokne and Ken Barker "Text Summarization Techniques: SVM versus Neural Networks" Proceedings of iiWAS2009.

[6] Massih R. Amini, Nicolas Usunier, and Patrick Gallinari, "Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms", Springer-Verlag Berlin Heidelberg 2005

[7] Michael Steinbach George Karypis Vipin Kumar "A Comparison of Document Clustering Techniques" In KDD workshop on Text Mining, 2002

[8] Neepa Shah, Sunita Mahajan "Document Clustering: A Detailed Review" International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5,October 2012

[9] Roma J, M S Bewoor, S H Patil, "Automation tool for Evaluation of the Quantity of NLP Based Text Summarization and Clustering Techniques By Quantitative and Qualitative Metrics" International Journal of Scientific & Engineering Research 2013

[10] Suneetha Manne Shaik, Mohammed Zaheer Pervez, Dr. S. Sameen Fatima "A Novel Automatic Text Summarization System with Feature Terms Identification" India Conference(INDICON), Annual IEEE 2011

[11] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998.

[12] Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda "Extracting Important Sentences with Support Vector Machines" ACM 2002.

[13] Vibekananda Dutta, Krishna Kumar Sharma Deepti Gahalot Performance Comparison of Hard and Soft Approaches for Document Clustering" International Journal Of Computer Applications March 2012