International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume3, No3., July 2012 – www.ijais.org

An Approach for Concept-based Automatic Multi-Document Summarization using Machine Learning

G.PadmaPriya
Assistant Professor
Department of Compuer Science and Engineering
K.S.R. College Of Engineering, K.S.R. Kalvi Nagar,
Tiruchengode, Tamilnadu, India.

K.Duraiswamy
Dean / Academic
Department of Compuer Science and Engineering
K.S.Rangasamy College Of Technology,
K.S.R. Kalvi Nagar,Tiruchengode, Tamilnadu, India.

ABSTRACT
Text Summarization is compressing the source text into a shorter version preserving its information content and overall meaning. It is very complicated for human beings to manually summarize large documents of text. Text summarization plays an important role in the area of natural language processing and text mining. Many approaches use statistics and machine learning techniques to extract sentences from documents. This paper presents an new approach for concept-based automatic multi-document summarization using machine learning.

General Terms
Data Mining, Text Mining.

Keywords
Multi-Document Summarization, Machine Learning.

1. INTRODUCTION
Natural Language Processing (NLP) is an area of research and application that analyze how computers are used for understanding and manipulating natural language text or speech to achieve the desired tasks. The goal of NLP researchers is to create an appropriate tools and techniques to make computer systems understand and manipulate natural languages by gathering the knowledge on how people understand and use languages [1]. There are more useful goals for NLP; most of them are related to the particular application for which it is being employed. The aim of the NLP system is to represent the exact meaning and purpose of the user's inquiry, which can be expressed in a usual language as if they were speaking to a reference librarian. Moreover, the contents of the documents that are being searched will be represented at all their levels of meaning so that a true match between need and reply can be found, despite how they are represented in their surface form [2]. Researchers mainly focus on techniques that have been developed in Information Retrieval [13] [15], while most try to influence both IR approaches and some features of NLP [14] [15]. In recent years there has been an explosion of on-line unstructured information in multiple languages, thus natural language processing technologies such as automatic text summarization have become increasingly important for the information retrieval applications.

Generally, text summarization [11, 12, 22-25] is the process of reducing a given text content into a shorter version by keeping its main content intact and thus conveying the actual desired meaning [3]. The summarization task can be classified as, either generic or query-oriented. A query-oriented summary presents the information which is salient to the given queries, whereas a generic summary provides an overall sense of the document's content. Document summarization can also be classified along two diverse dimensions such as abstract-based and extract-based. An extract-summary contains the sentences that are extracted from the document, whereas an abstract-summary may use words and phrases that do not present in the original document [5]. Abstraction involves recognizing a set of extracted sentences together constitute something new, that is not explicitly present in the source, and then substitute them in the summary with the new concept. Since the new material is not described in the original text, the system must have access to external information of some type such as, ontology or a knowledge base, and be able to perform combinatory inference [6]. Extraction method arranges the important sentences and comprises them into a summary. This method splits the document summarization task into three subtasks: 1) Ranking sentences based on their importance of being part in the summary, 2) Removing redundancy while extracting the salient sentences, 3) Systematizing the extracted sentences into a summary [7].

The automatic text summarization has two types of approaches: extraction and abstraction. The extraction method summarizes the text by using a limited number of sentences extracted from the original text, whereas the abstraction method produces a new, shorter text. Most research in automatic summarization has paid more attention on extraction, i.e., finding the most important clauses / sentences / paragraphs in texts [28] [29]. Extraction algorithms have a strong propensity to choose lengthy sentences from the text. The word frequency and distribution are often important, and are higher in lengthy sentences even when the sentence length is shortened. The resulting summary can be further reduced by shortening the extracted sentences and so that the gist of the sentence is preserved. Such summaries can surely reduce reading time of the user. Though, the extracting is a simple and robust method, it suffers from a number of problems. The one of the problem we focus is that the extracted sentences may be verbose and not achieve the goal of summarizing the document satisfactorily [27, 30].

Traditionally, summarization is made by humans, but due to the information overload problem, we must try to find the automatic way for summarization. Automatic text summarization is a well studied area and recently it has received a great attention due to vast amount of textual information available in electronic format [4]. Automatic Text Summarization is a method in which a computer summarizes

a text. A text is provided to the computer and it returns a concise and redundant-less extract of the original text. Summaries originate from two categories of text sources, a single document or a document sets [36]. Single document summarization can be defined as the process of creating a summary from a single text document. Multi-document summarization is the method of shortening, not just a single document, but a collection of related documents, into a single summary [37]. Commonly, a precise summary should be pertinent, short and articulate. In other words, the summary should meet the major concepts of the original document set, should be redundant-less and ordered [38]. These attributes are the basis of the generation process of the summary. The quality of summary is sensitive for those attributes relating to how the sentences are scored on the basis of the employed features. Consequently, the estimation of the efficacy of each attribute could result the mechanism to distinguish the attributes possessing high priority and low priority [39].

2. MOTIVATIONS FOR THE RESEARCH

Due to the increase in large volumes of information, companies can suffer from information overload and they may lose track in receiving the purposeful information. So, it is a time consuming task to scan through each of the long document [18]. A shorter version of the document that includes only the essential information is more beneficial for most information seekers. Therefore, implementing a text summarization system to create a summary needs more attention. In recent research, the text summarization has received a great deal of interest. It has the potential to summarize information automatically and present results to the end user in a compressed, yet complete form, which would help to solve the problem such as information overload. Further, for an Internet application, efficiency, even on large documents, is of substantial importance [19]. An efficient summarization method should satisfy the following three key requirements: • Diversity: A good document summary should be succinct and have as few redundant sentences as possible, i.e., two sentences giving same information should not be both present in the summary. In fact, employing diversity in summarization can efficiently decrease the redundancy among the sentences. • Coverage: The summary should enclose every important features of the document. By taking coverage into account, the loss of information in summarization can be reduced. • Balance: The summary should highlight the various features of the document in a balanced way. An unbalanced summary often cause great confusion about the general idea of the original document [17].

Text summarization focuses on the problem of selecting the most significant portions of the text, such as clauses, sentences, and paragraphs, and the problem of producing coherent summaries: in the context of multi-document summarization; in the context of revising single document extracts; in the context of headline generation; in the context of operations derived from an analysis of human written abstracts; in the context of sentence compression [20] [21]. The summarization analyze different levels of the text and determines which information in the text is relevant for a given summarization task. This determination of the importance of information in the source depends on a number of interacting factors such as the nature and type of the source text, the desired compression, and the information required by the application. The information needs must be able to address the user's needs effectively, and a summary of a text according to the user's interests and expertise must be considered, i.e., user focused summary vs. generic summary [21]. Automatic summarization is a technique, which enables a computer to summarize a larger text into a shorter form without any redundancy [26]. The benefits of automatic text summarization system's availability increase the need for existence of such system [23].

In general, a multi-document summary possesses some notable merits over a single-document summary. It offers a domain summary of a topic based on a document set representing identical information in several documents, distinct information in separate documents, and association between sections of information across various documents. It can enable the user to look in for more information on certain facets of interest, and look into the distinctive single-document summaries [37]. Most of the similar techniques employed in single-document summarization are also employed in multi-document summarization. There exist some notable disparities [13]: (1) The degree of redundancy contained in a group of topically-related articles is considerably greater than the redundancy degree within an article, since each article is appropriate to illustrate the most important point and also the required shared background. So, anti-redundancy methods play a vital role. (2) The compression ratio (that is the summary size with regard to the size of the document set) will considerably be lesser for a vast collection topically related documents than for single document summaries. When compression demands get intensified, summarization becomes challenging. (3) The co-reference problem in summarization possesses still bigger challenges for multi-document than for single-document summarization [31].

3. REVIEW OF RELATED WORKS

A handful of research works available in the literature deals about the automatic text summarization for single document and multiple documents. Here, we briefly reviewed some of the recent related works available in the text summarization field**.**

Dragomir R. Radev *et al.* [8] have presented a multi-document summarizer, MEAD, which created summaries by employing cluster centroids generated by topic detection and tracking system. It discussed two techniques, a centroid-based summarizer, and an evaluation scheme on the grounds of sentence utility and subsumption. The assessment was subjected to single and also multiple document summaries. In the end, they elaborated about two user studies that test the models of multi-document summarization.

Florian Boudin *et al.* [9] have presented a technique to topic-oriented multi-document summarization. It analyzed the efficacy of employing additional information about the document set all together, in addition to individual documents. The NEO-CORTEX, a multi-document summarization system on the basis of the available CORTEX system was furnished. Results are accounted for experiments with a document base created by the NIST DUC-2005 and DUC-2006 data. It was also showed that NEO-CORTEX was a competent system and realized better performance on topic-oriented multi-document summarization task.

Fu Lee Wang *et al*. [10] have presented a multi-document summarization system to obtain the critical information from terrorism incidents. News articles of a terrorism happening were arranged into a hierarchical tree structure. Fractal summarization model was used to produce a summary for all the news stories.

Yan Liu *et al.* [16] have proposed a document summarization framework via deep learning model, which has demonstrated distinguished extraction ability in document summarization. The framework consists of three parts of concepts extraction, summary generation, and reconstruction validation. A query-oriented extraction technique was proposed to concentrate information distributed in multiple documents to hidden units layer by layer. Then, the whole deep architecture was fine-turned by minimizing the information loss in reconstruction validation part. According to the concepts extracted from deep architecture, dynamic programming was used to seek most informative set of sentences as the summary. Experiments on three benchmark dataset demonstrated the effectiveness of the proposed framework and algorithms.

Shasha Xie and Yang Liu [32] have used a supervised learning approach for the summarization task and also used a classifier to determine whether to choose a sentence in the summary based on an affluent set of features. They have addressed two important problems related with the supervised classification approach. Firstly a diverse sampling technique has been proposed to handle the imbalanced data problem for the task in which the summary sentences are the minority class. Secondly a regression model has been used rather than binary classification for reframing the extractive summarization task in order to deal with human annotation disagreement problem.

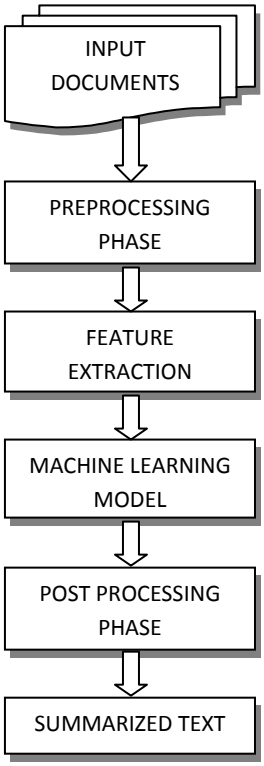
Allan Borra *et al.* [33] have aimed to develop a system that would be able to summarize a given document while still maintaining the reliability and saliency in the text. To achieve this, two main existing methods such as keyword extraction and discourse analysis based on Rhetorical Structure Theory (RST) have been included in ATS by the system architecture.

Rafeeq Al-Hashemi [34] has used an extractive technique to solve the problem with the idea of extracting the keywords, even if it does not exist explicitly within the text. The main role of their proposed project is the design of the keyword extraction subsystem which supports to select the more meaningful sentences to be in the summary. Their model contain four stages, mainly for the purpose of eliminating the stop words, extracting the keywords, ranking the sentences based on the keywords available in the sentences and finally for reducing the sentences using KFIDF measurement.

Gianluca Demartini *et al.* [35] have created an entity labeled corpus with temporal information beyond the TREC 2004 Novelty collection. They have developed and analyzed several features, and have demonstrated that an article's history could be used to enhance its summarization. The task of Entity Summarization (ES) is: given a query, a significant document and possibly a set of previous related documents (the history of the document), retrieve a set of entities which best summarizes the document.

4. PROPOSED SYSTEM

The primary intention of our research is to design and develop a system for concept-based automatic multi-document summarization using learning mechanism. With the intention of the framework available in [16], the learning mechanism and the concept model will be utilized in the proposed system. Initially, the input to the system will be multiple documents that have to be summarized. The documents utilized for text summarization is prepared by a set of preprocessing steps namely, sentence segmentation, tokenization, stop words removal and word stemming. Then, the concepts of every document are identified using the mutual information defined in the information retrieval. Then, the preprocessed documents will be given to feature extraction, which involves the identification of significance features like sentence length, term weight, thematic features, title features, key phrases, numerical data, and more. The proper learning mechanism will be used to learn the features extracted from the documents. Then, the most important sentences will be extracted as a summary from the multiple documents. Block diagram of the work is shown below



5. CONCLUSION

Automatic text summarization is an old challenge but the current research direction leans towards emerging trends in biomedicine, education domains, emails and blogs. This is due to the fact that there is information overload in these areas, especially on the World Wide Web. To study the cognitive aspect of text summarization we feel that use of machine learning is very helpful. Our approach will produce the considerable precision measures. In future we will try to implement this aspect to improve the performance of text summarization system.

6. ACKNOWLEDGMENT

The authors would like to thank all reviewers who have provided constructive feedback on this paper.

7. REFERENCES

[1] Gobinda G. Chowdhury, "Natural Language Processing", Annual Review of Information Science and Technology, Vol: 37, pp: 51–89, 2003.

[2] E D Liddy, "Natural Language Processing", In Encyclopedia of Library and Information Science, 2nd Edition, 2001.

[3] Inderjeet Mani, "Recent Developments in Text Summarization", In Proceedings of the tenth international conference on Information and knowledge management, ACM Press, pp: 529 - 531, 2001.

[4] Kaustubh Patil and Pavel Brazdil, "Sumgraph: Text Summarization Using Centrality in the Pathfinder Network", In IADIS International Journal on Computer Science and Information Systems, Vol.2, No. 1, pp: 18-32, 2007.

[5] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang and Zheng Chen, "Document Summarization using Conditional Random Fields", In Proceedings of IJCAI, 2007.

[6] Dragomir R. Radev, Eduard Hovy, Kathleen McKeown, "Introduction to the Special Issue on Summarization", In Computational Linguistics, Vol: 28, Issue 4, pp: 402, 2002.

[7] Jen-Yuan Yeh, Hao-Ren Ke and Wei-Pang Yanget, "iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network", Expert Systems with Applications, Vol: 35, pp: 1452, 2008.

[8] Dragomir R. Radev a, Hongyan Jing, Magorzata Sty and Daniel Tam, "Centroid-based summarization of multiple documents", Information Processing and Management, vol. 40, no.6, pp. 919–938, 2004.

[9] Florian Boudin and Juan Manuel Torres Moreno, " NEO-CORTEX: A Performant User-Oriented Multi-Document Summarization System", Lecture Notes in Computer Science", Springer, Vol 4394, pp.551-562, May 2007.

[10] Fu Lee Wang, Christopher C. Yang and Xiaodong Shi," Multi-document Summarization for Terrorism Information Extraction", Lecture Notes in Computer Science", Springer, vol. 3975, May 2006.

[11] Aaron Harnly, Ani Nenkova, Rebecca Passonneau and Owen Rambow, "Automation of Summary Evaluation by the Pyramid Method", In Proceedings of the Conference of Recent Advances in Natural Language Processing, pp: 226, 2005.

[12] Rachit Arora and Balaraman Ravindran, "Latent Dirichlet Allocation Based Multi-Document Summarization", In Proceedings of the second workshop on Analytics for noisy unstructured text data, pp:91-97, 2008.

[13] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, "Multi-document summarization by sentence extraction", ANLP/NAACL Workshop, pp: 40–48, 2000.

[14] E. Hovy and C. Lin., "Automated text summarization in SUMMARIST", In Advances in Automatic Text Summarization, 1999.

[15] James Allan, Rahul Gupta, and Vikas Khandelwal, "Temporal Summaries of News Topics", 2001.

[16] Yan Liu, Sheng-hua Zhong, Wen-jie Li, "Query-oriented Unsupervised Multi-document Summarization via Deep Learning", Under review in Journal of Neural Networks (NN).

[17] Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, Yong Yu, "Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning", In www 2009 madrid, pp: 71-72, 2009.

[18] L. H. Chong, and Y. Y. Chen, "Text Summarization for Oil and Gas News Article", 2009.

[19] H. Gregory Silber, Kathleen F. McCoy, "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization", Association for Computational Linguistics, Vol: 28, 2002.

[20] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., Eskin, E., "Towards multi-document summarization by reformulation: Progreess and prospects", In Proceedings of the Sixteenth National Conference on Artificial Intelligence, pp: 293, 1999.

[21] Yllias Chali, "Generic and Query-Based Text Summarization Using Lexical Cohesion", Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Springer-Verlag London, pp: 293-302, 2002.

[22] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", Second IEEE International conference on intelligent systems, pp: 40-45, 2004.

[23] M. S. Binwahlan, N. Salim, L. Suanmali, "Intelligent Model for Automatic Text Summarization", Information Technology Journal, pp: 1249-1255, 2009.

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

[25] H. Edmundson, "New methods in automatic extracting", Journal of the Association for Computing Machinery, Vol: 16, No. 2, pp: 264-285, 1969.

[26] Hassel M., "Resource Lean and Portable Automatic Text Summarization", PhD thesis, School of Computer Science and Communication, 2007.

[27] Michel Gagnon, Lyne Da Sylva, "Text Summarization by Sentence Extraction and Syntactic Pruning", In Proceedings of Computational Linguistics in the North East, 2005.

[28] Kevin Knight, Daniel Marcu, "Summarization beyond sentence extraction:A probabilistic approach to sentence compression", In Artificial Intelligence, Vol: 139, Issue 1, pp: 91–107, 2002.

[29] I. Mani, M. Maybury, "Advances in Automatic Text Summarization", MIT Press, 1999.

[30] H Jing, K McKeown, "The decomposition of human-written summary sentences", In 22nd International Conference on Research and Development in Information Retrieval, pp: 129-136, 1999.

[31] Breck Baldwin and Thomas S. Morton, "Dynamic coreference-based summarization", in Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada, Spain, June 1998.

[32] Shasha Xie, Yang Liu, "Improving supervised learning for meeting summarization using sampling and regression", In Computer Speech and Language, Vol: 24, Issue 3, 2009.

[33] Allan Borra, Almira Mae Diola, Joan Tiffany T. Ong Lopez, Phoebus Ferdiel Torralba, Sherwin So, "Using Rhetorical Structure Theory in Automatic Text Summarization for Marcu-Authored Documents", In titaniaadduduph, 2010.

[34] Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSES)Using Extracted Keywords", International Arab Journal of e-Technology, Vol. 1, No. 4, pp: 164-168, 2010.

[35] Gianluca Demartini, Malik Muhammad Saad Missen, Hugo Zaragoza, "Entity Summarization of News Articles", In SIGIR, pp: 795-796, 2010.

[36] Liang Zhou, Miruna Ticrea and Eduard Hovy, "Multi-document Biography Summarization", in Proceedings of Empirical Methods in Natural Language Processing, 2004.

[37] Shiyan Ou, Christopher S.G. Khoo and Dion H. Goh, "Design and development of a concept-based multidocument summarization system for research abstracts", Journal of Information Science, vol. 34 , no. 3, pp. 308-326 , June 2008.

[38] You Ouyang, Wenji Li and Qin Lu, "An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation", in proceedings of the ACL-IJCNLP, singapore, pp. 109–112, 2009.

[39] Mohammed Salem Binwahlan, Naomie Salim and Ladda Suanmali, "Swarm Based Features Selection for Text Summarization", IJCSNS International Journal of Computer Science and Network Security, vol. 9, no.1, January 2009.