A Modified Technique for Modern Multi-Document Summarization

Mr. Vijay Sonawane[1] Prof. Amit Mishra[2] Dr. Shiv Sahu[3]
[1,2,3]Department of Information Technology
[1,2,3]Technocrats Institute of Technology, Anand nagar, Bhopal, India

Abstract— The text summarization is the need of the hour. as the information is rapidly growing on the internet. The problem is that when you search for any information the internet provides more information than the required. So there is a need to summarize the result. Also there is a need to maintain the quality of information in the summary. In this paper, we have proposed a novel modified technique for the multi document summarization.

Key words: summarization, multi-document summarization

I. INTRODUCTION

The text summarization [1] has become an important, integral and timely tool for assisting and interpreting text information in modern information era. It is a very complex process & It is very difficult for human beings to manually summarize large documents of text. There are tons of material available on the internet. Generally the information provided by the internet is more than the required. So there are basic problems identified as: locating the relevant documents from thousand number of documents available, and also maintaining the quantity of relevant information. The primary goal of automatic text summarization is to convert the source text into a shorter version by preserving its information content and overall meaning.

One way to employ a summary [3] is in an indicative way as a pointer to some parts of the original document or in informative way to cover all relevant information of extractive the text. The common advantage of both methods of using a summary is its reduced reading time. The common characteristic of a good automatic summary system is that summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. The tool must also search for headings and other markers of subtopics in order to identify the key points of a document. One example of such tool is Microsoft Word's AutoSummarize function.

II. CLASSIFICATION OF TEXT SUMMARIZATION METHODS

The multi-document summarization text summarization methods can be classified into extractive and abstractive summarization. The extractive summarization method consists of selecting important sentences, and paragraphs from the original document and concatenating them into shorter form. The criteria used for deciding the importance of sentences is based on statistical and linguistic features of sentences.

One more common way, the abstractive summarization [5][6] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. Abstractive summarization 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.

Another class is extractive summaries [2]. These are formulated by extracting key text segments from the text. It is based on statistical analysis of individual or mixed surface level features such as word frequency to locate the sentences to be extracted. In this method, the most important content is treated as the most frequent or the most favorably positioned content. This approach thus avoids any efforts on deep text understanding. Such methods are conceptually simple and easy to implement.

The overall process of extractive text summarization process [4] can be divided into two steps: Pre Processing step and Processing step.

The Pre Processing step is structured representation of the original text. This step includes the Sentences boundary identification. The general meaning of sentence boundary is identified with presence of dot at the end of sentence. Also includes stop word elimination. It means that the common words with no semantics and which do not aggregate relevant information to the task are eliminated. Then it includes stemming. The goal of stemming is to obtain the stem or radix of each word. It emphasizes its semantics.

In second step or the Processing step, all the features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Then in subsequent step the final score of each sentence is determined using feature weight equation. At the end, the top ranked sentences are selected for final summary.

III. LITERATURE SURVEY

Multi document summarization became more interested by the mid-1990s. The summary from multi document must include the important ideas in each document. Also it should compare ideas across document and reduce the size of each document and ordering in new sentence.

Radev and McKeown [7] developed SUMMONS to generate summaries of multiple documents on the same or related events. They do so by presenting similarities and differences, the contradictions, the generalizations among sources of information from realized as English sentences.

They improved their SUMMONS in 1998 [10], they combined it into a conceptual representation of the summary which selects information from underlying knowledge base. The lower rating is given to structured conceptual representation of the summary. In such case where the information appears in only one article is and information that is synthesized from multiple articles is rated more highly.

Authors Kathleen R. McKeown et al [12] presented Multi Gen and DEMS for Columbia multi-document summarization system built on the observation that depending on the intended purpose of the summary and on the types of document summarized. They focused on the summarization of sets of documents that all describe the same event or news. The method used an enhanced version of Multi Gen to summarize the document. Technique used alternative system DEMS known as Dissimilarity Engine for Multidocument Summarization for biographical documents. The input articles are transformed into a uniform XML format in the processing step. Router components of the system determined the type of each input document set and direct the input texts to the summarizers.

In 2004, Fukumoto [9] proposed a summarization system which automatically classified type of document set and summarized a document set with its appropriate summarization mechanism. The system classified a document set into three types, a. a series of events, b. a set of the same events and c. related events, by using information of high frequency nouns and named entity. All the unnecessary parts are deleted after summarized each document and generated multi-document summary. This technique used single document summarization mechanism for each document of a document set and removed similar parts between summarized documents for generation of a target summary. Author applied a TF based sentence extraction for single document summarization and used of single document summarization for multi-document summarization. But the mechanism of document set classification does not work well in the evaluation because their current implementation has some system bugs in classification mechanism.

In 2005, Yan-Min [11] used lexical chain for multi document summarization in Chinese document based on How net knowledge database. The algorithm starts from pre-process the text. In this step it removes redundant similarities and remain differences in information content among multiple documents, then it constructs lexical chains and identifies strong chains. After that the significant sentences are extracted from each document and ordered sentences, then it perform recognition and removal of redundant information. The summary is generated in chronological order at the end. The anaphora resolution technology is applied to improve the fluency of the summary.

In 2008, the authors Judith D. Schlesinger et al [8] proposed CLASSY system. The CLASSY text summarization is an automatic multi-lingual document summarization. The CLASSY system architecture to summarize document. CLASSY stands for Clustering Linguistics And Statistics for Summarization Yield, It is an automatic summarization system that uses linguistic trimming and statistical methods to generate generic or topic driven summaries for single documents or clusters of documents. It used trimming rules to shorten sentences, to identify sentences, to select sentences and to organize the selected sentences for the final summary. The primary objective of this research is to generate a multi-lingual summarization document based on summaries document from Machine Translate document. This architecture consists of five steps: a. document preparation, b. sentence trimming, c. sentence scoring, d. redundancy reduction, and e. sentence ordering. This system will generate very good summaries when using signature term that computed from English document and machine translated version of Arabic document.

IV. PROPOSED APPROACH

In this paper, we have proposed a variant of the existing STARLET technique. It is a novel multi document summarization technique. This technique combines the advantages of both abstractive & the extractive method. This method takes a set of reviews as input & then each input is labeled aspect rating.

It generates abstractive or extractive reviews as output. These reviews are informative, concise, less redundant, readable, and target oriented.

This model is shown below in figure:

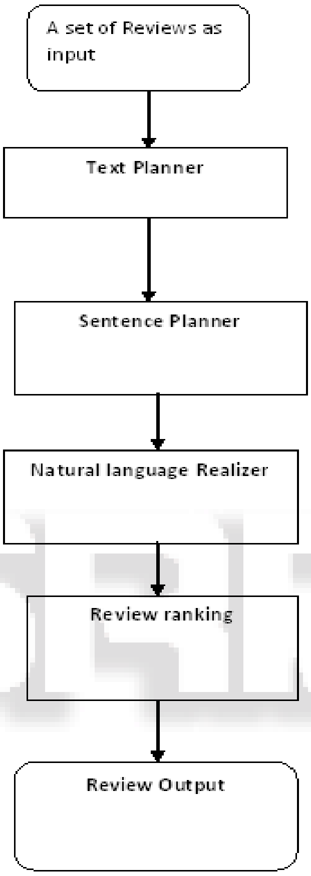

Fig.1: System model

V. CONCLUSION

In this paper, we have elaborated the concept of the multi document summarization. We have also given a classification of the text classification. The literature review of the modern multi document summarization technique is given. The working, the merits and the demerits of each text summarization technique is discussed. This paper will help researchers of the common field in understanding the concept and the getting the review of some common text summarization methods.

REFERENCES

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.112, ISBN 978-80-227-2827-0, FIIT STBrarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[2] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008. Springer Berlin / Heidelberg, Vol 2507/2002, 205-215, 2002.

[3] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005.

[4] Vishal Gupta, G.Sl Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 60-76, AUGUST 2009.

[5] G Erkan and Dragomir R. Radev, "LexRank: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004.

[6] Udo Hahn and Martin Romacker, "The SYNDIKATE text Knowledge base generator",Proceedings of the first International conference on Human language technology researchAssociation for Computational Linguistics , ACM, Morristown, NJ, USA , 2001.

[7] Dragomir R. Radev and Kathleen R. McKeown. 1995. Generating summaries of multiple news articles. In proceeding of SIGIR'95, Seattle, Washington. 74–82.

[8] Judith D. Schlesinger, Dianne P. O'Leary, and John M. Conroy. 2008. Arabic/English Multi-document Summarization with CLASSY—The Past and the Future. In Computational Linguistics and Intelligent Text Processing, 9th International Conference, Haifa, Israel, February 17-23, 2008, Proceedings. Springer. 568–581.

[9] Jun ichi Fukumoto. 2004. Multi-Summarization using document set type classification. In Proceedings of NTCIR4, Tokyo.

[10] Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple online sources. Computational Linguis-tics, 4:469–500.

[11] Yan-Min C., Xiao-Long W. and Bing-Quan L., 2005. MultiDocument Summarization Based on Lexical Chains. In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou. IEEE, 1937-1942.

[12] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou,Simone Teufel, M. Yen Kan, andBarry Schiffman. 2001. Columbia multi-document summarization: Approach and evaluation. In Donna Harman and Daniel Marcu, editors, Proceedings of the First Document Understanding Conference (DUC '01), New Orleans, Louisiana, USA. 1-21