International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Comparison of Multi Document Summarization Techniques - A Survey

Heena Kishor Patil[1], Kumud Wasnik[2]

[1]M. Tech Computer Science & Technology Student, UMIT, SNDT University, Mumbai, India

[2]Professor, Computer Science & Technology Department, UMIT, SNDT University, Mumbai, India

Abstract: ***Huge amount of information is present on the World Wide Web and a large amount is being added to it frequently. It is a tedious task for the user to go through all these documents as the number of documents available on a topic will range from tens to thousands. When we see the large amount of information on the web about a particular topic, we always feel that there is a need of some automated tool. Multi-Document Summarization helps at extraction from a set of documents written about same topic and helps to familiarize themselves with information content in large set of documents.***

Keywords: Summarization, Abstractive, Extractive, Feature Based, Cluster Based, Query Dependent

1. Introduction

Currently, the World Wide Web is the largest source of information. There is huge amount of data available on the web with structured and unstructured form and it is difficult task to read all data within less time. Hence we need a system that automatically retrieves and summarize the documents as per user need within short time. Document Summarization is one of the feasible solutions to this problem. Summarization extract relevant portion of information from a set of documents about same topic and represent them in coherent order. Summaries can be defines using Single Document Summarization and Multiple Document Summarization. The summary which is created from single document is called as Single Document Summarization whereas Multiple Document Summarization is an automatic process for the extraction of information from multiple text documents.

The main goal of summarization is to create Summary which covers all the major aspects of original document without missing relevancy between the summary documents within short time. Summary generation can be broadly divided as abstractive and extractive. In abstractive summary generation, the abstract of the document is generated and it will not have exact sentences as present in the document. In extractive summary generation, important sentences are extracted from the document. The generated summary contains all such extracted sentences arranged in a meaningful order.

This survey focuses on advantages of other proposed multi-document summarization over the existing multi - document summarization strategy. The main aim of Multi Document Summarization has been also explained. We examine the remarkable methodologies of multi document summarization and present it with related work literature. The rest of the study is sorted out as follows: First we display the review on four multi document summarization approaches to be specific the feature based technique, cluster based strategy, graph based system. At that point we point the proposed multi document summarization strategy; i.e. Query dependent increment multi document using clusters. At last we end with conclusion.

2. Literature Review

A number of exploration studies shows different sorts of methodologies and accessible systems for multi document summarization. In this study we guide our focus remarkably on techniques to do multi document summarization. Our survey talks on primary summarization strategy and research study from related literary works [1][2][3][4].

A. Title & Paragraph (Feature) Based Summarization:

Table 1: Title and Paragraph(Feature) based summarization

ID	Cluster (Title)	Cluster (Paragraph)	Category	Total Documents
1	UP Polls 2017	Uttar Pradesh's political heavy-weights who have money,	Politics	10
2	CBSE Board Exam	To sort things out and to help all CBSE board	Education	10
3	Notes Ban Created	Prime Minister Narendra Modi	Economy	10

Research object is used for this method is only news documents because news documents typically have a title which is relevant to the whole body of the document. The method used in this research is hybrid abstractive-extractive summarization technique which the combination of WordNet based text summarization (abstractive technique) and title word based text summarization (extractive technique).

B. Summary System and Human Based Document Summarization

This technique presents a novel approach to generate an abstractive summary from extractive summary using WordNet ontology. An experimental result shows the generated summary in well-compressed, grammatically correct and human readable format. In this method, system generated summary compared with original human generated document and finally meaningful human readable

summary generated by using abstractive summary generation

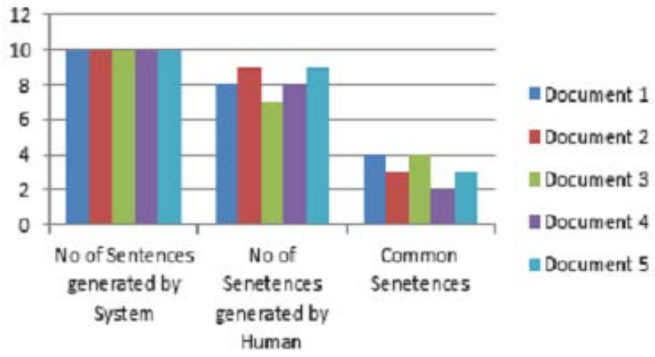

Figure 1: Comparison between human generated summary and summary generated by system [2]

C. Sentence (Clustering) Based Summarization

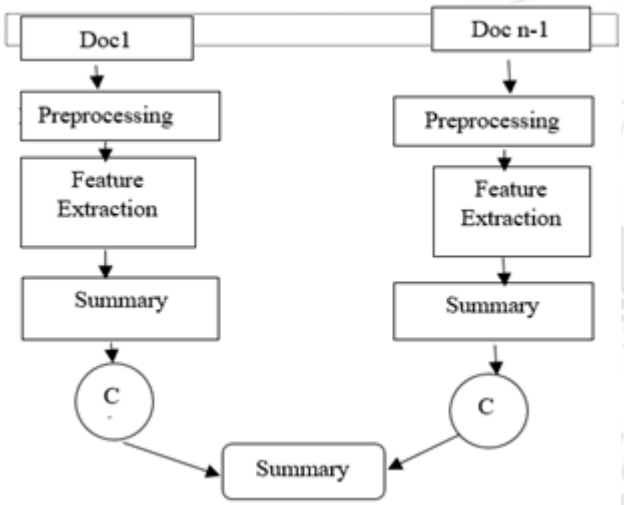

Figure 2: Comparison between human generated summary and summary generated by system

As the name indicate it generates summary on the basis on sentence. Sentence from single document summaries are clustered and top most clustered from each document used for creating multi document summarization. Both syntactic and semantic similarity between sentences is used for clustering. A number of semantic similarity measures based on literature concepts. The syntactic similarity used in this method is based on word order, which can be replaced with other structural similarity measures.

D. Graph Based Summarization

For graph based multi document summarization where existence of an edge between a pair of sentences is determined based on how much two sentences are similar to each other. Since the documents may contain redundant information, the performance of a multi document summarization system mainly depends on the sentence similarity measure used for removing redundant sentences from the summary.

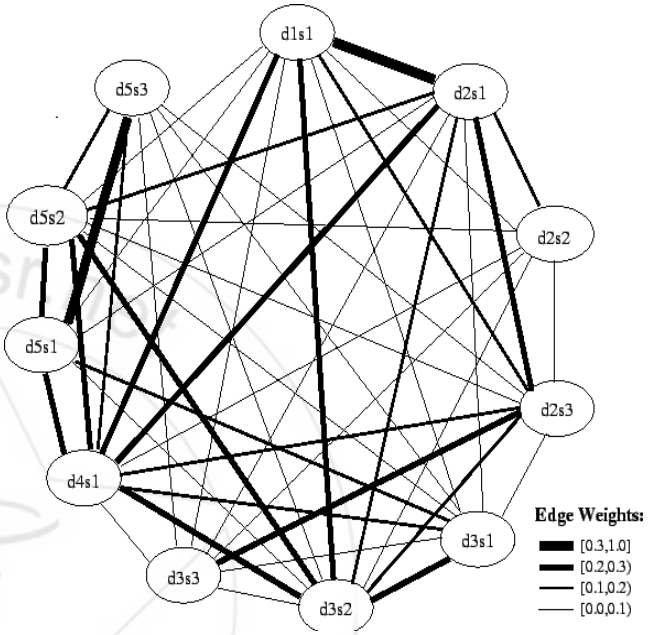

Figure 4: Graph Based Summarization

E. Proposed System- Query Dependent Increment Multi Document Summarization Using Clusters

Query dependent systems generates extractive summary which is influenced by the query. The query is also analyzed semantically and provided as an input to the system. The summary generation process is guided by the information contained in the query. The incremental approach uses cluster based and feature based summarization for summary generation and on the basis of each word and sentence weight we can show edge weight graph representation also.

Table 2: Comparison of Different Summarization Techniques

Techniques	Summary Trigger	Application	Summarization approach	Dependency	Summary Construction Method	Limitations
Title & Paragraph(Feature) Based Summarization	Generic	News Paper	Feature Based	Paragraph Title	Extractive	Fails to calculate summary if information has no titles and paragraph.
Summary System and Human Based Document Summarization	Generic	Scientific Theory	Evaluator Based	Evaluator self-generated Summary	Abstractive	Fails to calculate summary if evaluator summary is not Provided.
Sentence (Clustering) Based Summarization	Generic	Data Mining	Cluster Based	Semantic and symmetric sentence	Extractive	It may provide summary with same meaning although it has used different words
Graph Based Summarization	Generic	Opinion Polls	Graph Based	Accurate weight of sentence	Extractive	Similarity Computation may be slow down the summarization process.
Proposed System- Query	Query	News, Social	Feature Based,	Query weight and	Extractive	Cannot add multiple

Dependent Increment Multi Document Summarization Using Clusters	Dependent	media, Data mining etc.	Graph Based, Clustered Based	document clustering		document at same time .it may crash down the system.

3. Comparative Study

In this section, we compare the existing summarization techniques like graph based, cluster based and feature based multi document summarization method with the proposed system i.e.Query dependent increment multi document using clusters. Table II will help to understand the comparative studies between summarization techniques.

4. Problem Statement

WWW World Wide Web is the largest source of information. The huge amount of information is available on internet and it is increasing day by day. The need of text summarization has recently increased due to the proliferation of information on the Internet, With the availability and speed of internet. Information search from online documents has been eased down to users finger tips. However, it is not easy for users to manually summarize those large online documents. Hence the query dependent summary approach generates a shorter version of document content and overall meaning in incremental way and this process takes take will less time.

5. Conclusion

This study gives a brief overview on multi document summarization techniques. Four sorts of methodologies have been talked about, in particular the feature based system, cluster based strategy, graph based system and evaluator based strategy. It gives the idea about each of these techniques has its own particular favorable circumstances towards multi document summarization. In the meantime, there are a few limitations alternately restrictions relating to those techniques. For future work, a query dependent methodology taking into account to create a summary which is appropriate for an informative type summarization era. We conviction that the query based method technique can reduce a percentage of the previously stated limits.

References

[1] Glorian Yapinus, Alva Erwin, Maulahikmah Galinium and Wahyu Muliady," Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive-Extractive Summarization Technique", 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia.
[2] Harsha Dave and Shree jaswal, "Multiple Text Document Summarization System using Hybrid Summarization Technique" in 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015) Dehradun, India, 4-5 September 2015.
[3] Virendra Kumar Gupta and Tanveer J. Siddiqui, "Multi-Document Summarization Using Sentence Clustering", IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction, Kharagpur, India, December 27-29, 2012.
[4] Linhong Zhu, Sheng Gao, Sinno Jialin Pan, HuizhouLi, Dingxiong Deng and Cyrus Shahabi," Graph-based Informative-Sentence Selection for Opinion Summarization"; 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.
[5] Atif Khan, Naomie Salim and Haleem Farman; "Clustered genetic semantic graph approach for multi-document abstractive summarization"; Intelligent Systems Engineering (ICISE), 2016 International Conference on.
[6] Kamal Sarkar, Khushbu Saraf and Avishikta; "Improving Graph Based Multidocument Text Summarization Using an Enhanced Sentence Similarity Measure";2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS).