Study on Multi Document Summarization by Machine Learning Technique for Clustered Documents

Sweta Kasundra, Prof. Devangi L. Kotak

Computer Engineering Department
V.V.P. Engineering Collage, Rajkot, India.

***Abstract* - This paper discusses the development of multi document summarization by using different approach like abstractive-extractive summarization approach. Multi document summarization is a technology that use to summarize multiple documents and make its summary. A particular challenge for multi-document summarization is that there is an information stored in different documents. In this paper we discuss different approaches used like LSA, LDA, LDA-SVD, Semantic Graph approach etc.**

***Index Terms* - Text Summarization, Multi document summarization, Abstractive Technique, Extractive Technique.**

I. Introduction

Nowadays, we are facing the problem of information overload. The ever-increasing huge amount of textual documents needs to be analyzed and summarized efficiently and effectively. Thus, the demand for extensive study in the automatic multi-document summarization is growing rapidly as well. Multi-document summarization summarizes information from multiple documents which share similar topics. It can help users quickly sift through large text data collections, catch the most relative or important information. Multi-document summarization has been applied in a wide range of domains, from the traditional such as newswire or scientific articles summarization, to novel domains such as literary text, patents, or blog post summarization and twitters analysis.

Multi-document summarization methods can be either extractive or abstractive. An abstractive summarization attempts to gain an understanding of the main concepts in a document [6]. On the other hand, extractive summarization attempts to extract parts of the original document such as important sentences and paragraphs in order to concatenate them into a summary. Most of the studies have focused on multi-document extractive summarization using techniques of sentence extraction, statistical analysis, discourse structures and various machine learning techniques. Where Abstractive summarization requires natural language processing techniques such as semantic representation, natural language generation, and compression techniques.

II. Literature Survey

In Paper [1] , They proposes the use of lexical similarity across different documents in order to improve a topic segmentation task. Given a set of topically related documents, the segmentation process is carried out using a Bayesian framework. By using similar sentences from different documents more accurate segment likelihood estimations are obtained. The proposed approach was tested in an educational domain where a set of learning materials from different media sources needed to be segmented so that students could browse through them more efficiently. Initial results show that the proposed method does afford better segmentation compared to one of the present state of the art algorithms, a Bayesian baseline approach that segments the documents individually.

In Paper [2] , They discusses the development of multi document summarization for Indonesian documents by using hybrid abstractive-extractive summarization approach. Multi document summarization is a technology that able to summarize multiple documents and present them in one summary. The method used in this research, hybrid abstractive-extractive summarization technique, is a summarization technique that is the combination of WordNet based text summarization (abstractive technique) and title word based text summarization (extractive technique). After an experiment with LSA as the comparison method, this research method successfully generated a well-compressed and readable summary with a fast processing time.

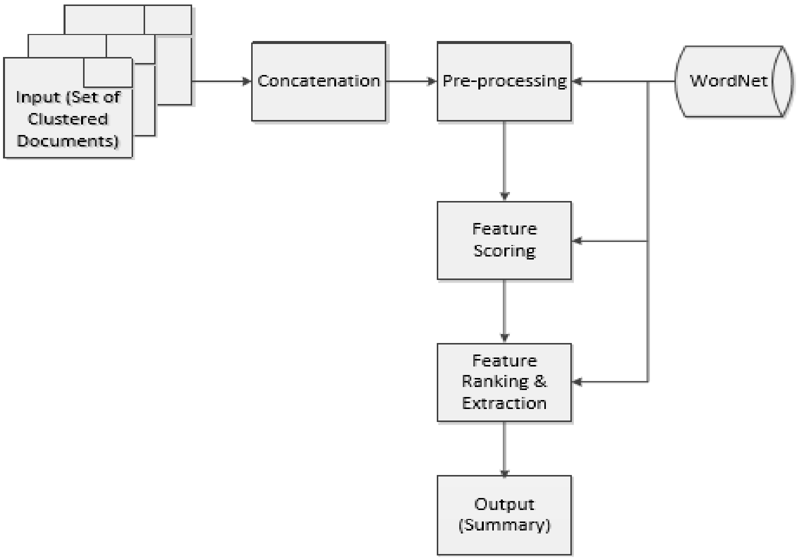

Fig-1 Methodology for paper[2]

In Paper [3] , In this paper, they proposed a novel pattern-based topic model (PBTMSum) for the task of the multi-document summarization. PBTMSum combining pattern mining techniques with LDA topic modelling could generate discriminative and semantic rich representations for topics and documents so that the most representative and non-redundant sentences can be selected to form a succinct and informative summary. Extensive experiments are conducted on the data of document understanding conference (DUC) 2007. The results prove the effectiveness and efficiency of our proposed approach.

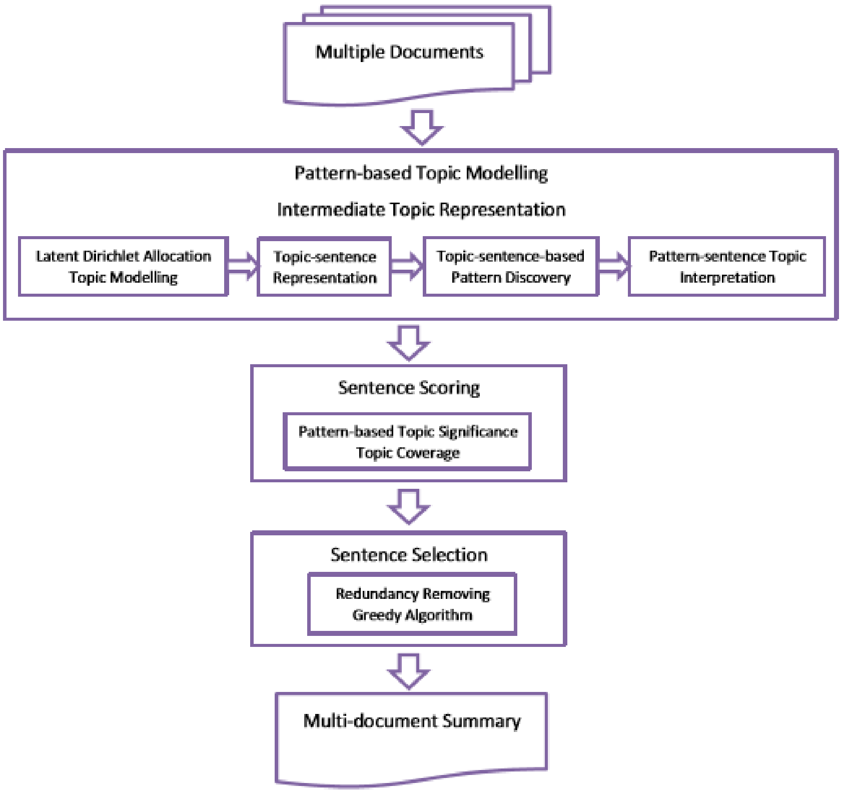
Multiple Documents
Pattern-based Topic Modelling
Intermediate Topic Representation
Latent Dirichlet Allocation Topic Modelling
Topic-sentence Representation
Topic-sentence-based Pattern Discovery
Pattern-sentence Topic Interpretation
Sentence Scoring
Pattern-based Topic Significance
Topic Coverage
Sentence Selection
Redundancy Removing
Greedy Algorithm
Multi-document Summary

Fig-2 Methodology for Paper[3]

In paper [4] , They use Latent Dirichlet Allocation can break down these documents into different topics or events. However to reduce the common information content the sentences of the summary need to be orthogonal to each other since orthogonal vectors have the lowest possible similarity and correlation between them. Singular Value Decomposition is used to get the orthogonal representations of vectors and representing sentences as vectors, we can get the sentences that are orthogonal to each other in the LDA mixture model weighted term domain. Thus using LDA we find the different topics in the document sand using SVD. we find the sentences that best represent these topics. Finally we present the evaluation of the algorithms on the DUC

2002Corpusmulti-documentsummarizationtasksusingthe ROUGE evaluator to evaluate the summaries. Compared to DUC 2002 winners, our algorithms gave significantly betterROUGE-1recallmeasures.

In paper [5] , They introduces a clustered genetic semantic graph approach for multi-document abstractive summarization. The semantic graph from the document set is constructed in such a way that the graph vertices represent the predicate argument structures (PASs), extracted automatically by employing semantic role labeling (SRL); and the edges of graph correspond to semantic similarity weight determined from PAS-to-PAS semantic similarity, and PAS-to-document relationship. The PAS-to document relationship is expressed by different features, weighted and optimized by genetic algorithm. The salient graph nodes (PASs) are ranked based on modified weighted graph based ranking algorithm. The clustering algorithm is performed to eliminate redundancy in such a way that representative PAS with the highest salience score from each cluster is chosen, and fed to language generation to generate summary sentences. Experiment of this study is performed using DUC-2002, a standard corpus for text summarization. Experimental results indicate that the proposed approach outperforms other summarization systems.

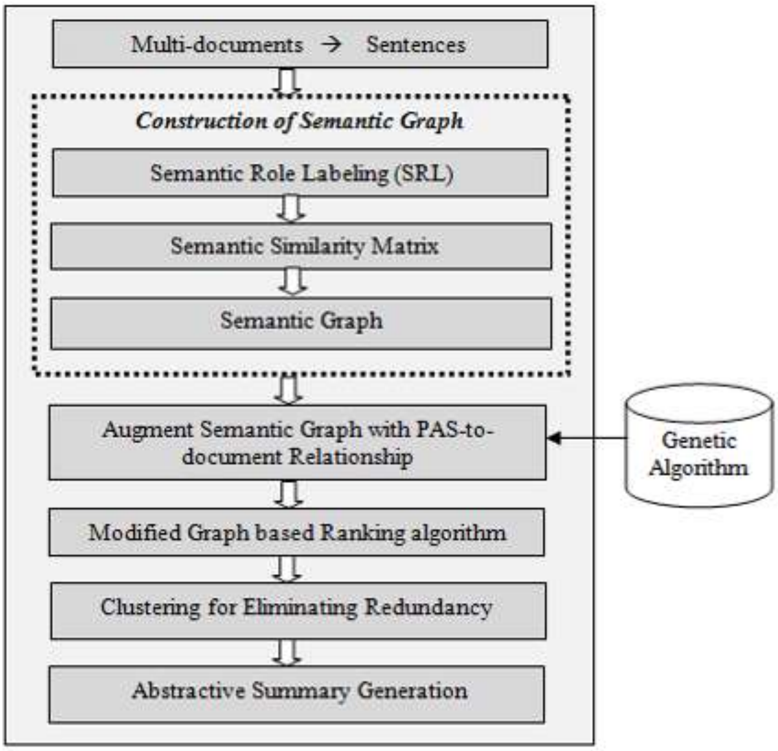
Multi-documents → Sentences
Construction of Semantic Graph
Semantic Role Labeling (SRL)
Semantic Similarity Matrix
Semantic Graph
Augment Semantic Graph with PAS-to-document Relationship
Genetic Algorithm
Modified Graph based Ranking algorithm
Clustering for Eliminating Redundancy
Abstractive Summary Generation

Fig-3 Methodology for Paper[5]

In Paper[6] , Many work for text summarization has been represented in academia. In the world of internet vast amount of information is accessed by the user but still user wants this information in short and meaningful way. Summarization has the answer to this requirement. Summarization can be applied to a single document or multiple documents. When the summarization is applied to numerous documents, the summarization is called as multi-document summarization. There are mainly two types of summarization 1.Extractive summarization 2.Abstractive summarization. In Extractive summarization, important sentences are identified and only those sentences are included in summary, desired summary length is obtained by use of compression ratio. In case of abstractive summarization, according to scoring criteria recognize relevant sentences and process these sentences so as the sentences can be included in summary. Abstractive summarization includes deep understanding of natural language and it is also based on compression. Mostly automatic summarization deals with extractive type of summarization.

They proposed system which is based on the cross-document structure theory. Cross-document structure theory generates summary from documents which are relevant to each other. Figure below shows general architecture of the proposed approach. This proposed architecture focuses on two main things CST relation identification and sentence scoring. While considering multiple documents as input to produce summary, we need to identify the type of CST (cross document structure theory) relationship that is present among the documents from same domain. Dr. D. R. Radev discussed [6] 24 types of CST relations that present among the documents with same domain. Multi document graph can be drawn at the four levels, word level, phrase level, sentence level, and document level. Cross document relationships are identified at all this levels with the help of CST relations. Proposed technique makes use of some of the CST relations. First step in proposed system is preprocessing the documents after that features are extracted. And depending on this by making use of some of CST relations cross-document relationships are identified by making use of dictionary. In sentence scoring, score of sentence is calculated using scoring model. Final summary is generated by defining threshold value for score of sentence.

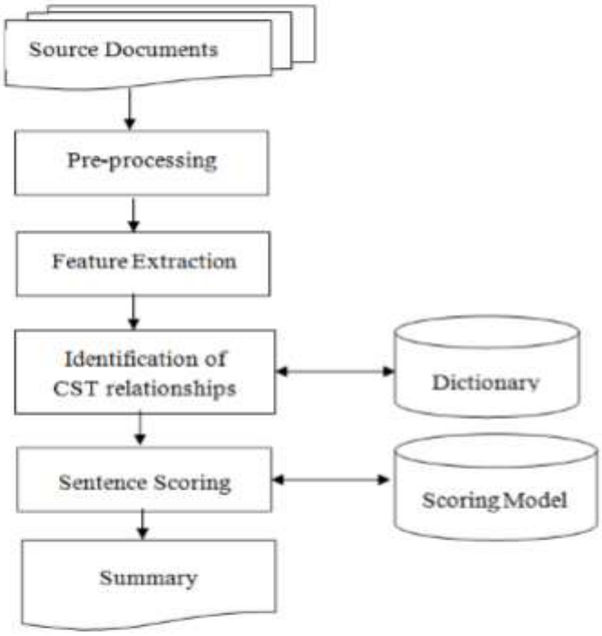

Fig-4 Methodology for paper[6]

Now we see the comparison of all methodologies used in 6 papers. Table-1given below shown the methodology, algorithm used and future work of all literatures.

Table-1 Comparison Table

Paper	Literature	Algorithm Used	Issues	Future Work
[1]	Multi-Document Topic Segmentation using Bayesian Estimation	Bayesian Methods	Enhance the topic segmentation of an individual document by incorporating information from other documents.	Develop more sophisticated modeling techniques. Capture how different subtopics are reused across documents.
[2]	Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive Extractive Summarization Technique	Latent Semantic Analysis (LSA)	To ascertain whether the two current text summarization techniques, abstractive and extractive, can be combined together in order to generate a fast generated, well-compressed, and readable summary.	Readability level can be increased if the natural language processing techniques are used. Use Automatic clustering instead of using manually cluster documents. The summarization accuracy can be increased if the categories and their keywords that generated manually in this research are replaced with expanded categories with keywords that generated statistically.
[3]	Mining Topical Relevant Patterns for Multi-document Summarization.	Latent Dirichlet Allocation (LDA)	To exploit pattern based topic model to automatically generate discriminative and semantic rich representations for multi-document summarization.	Optimize the algorithm for redundancy removing to enhance the performance of the PBTMSum model. Co-reference and sentence ordering into consideration to develop more accurate solution for multi-document summarization with pattern-based topic modelling. Use different optimizing methods in LDA topics modelling phase.

[4]	Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization	LDA, SVD	Using LDA we can break down the documents into topics using a mixture model. Thus each Document is a mixture model over the Topics. The SVD derives the latent semantic structure from the sentence represented by matrix.	Replacing LDA with other topic models.
[5]	Clustered Genetic Semantic Graph Approach for Multi-document Abstractive Summarization	Graph based ranking algorithm	Propose a clustered genetic semantic graph approach for multi-document abstractive summarization. Propose modified weighted graph based ranking algorithm to take into account the PAS-to-PAS semantic similarity and PAS-to-document relationship.	Explore Cross-Document Structural Theory (CST) relations for multi-document abstractive summarization and examine their impact on summarization.
[6]	Multi Document Summarization: Approaches and Future Scope	CST (Cross-document structure theory)	The source documents and ranking of sentence is done by using scoring model.	Manually clustering is used. Instead of that use automatic clustering.

III. CONCLUSION

In this paper we do the survey of different literatures and understand the different methodologies used for multi document summarization. We see the different algorithms like LSA, LDA, LDA-SVD, Graph base ranking algorithm etc. We also see the proposed system published in those literatures for Abstractive approach and Extractive approach. Thus, we can use different methodology to increase the effectiveness and reduce time.

REFERENCES

[1] Pedro Mota, Maxine Eskenazi, Luisa Coheur, “Multi-Document Topic Segmentation using Bayesian Estimation”, Tenth International Conference on Semantic Computing, IEEE-2016.

[2] Glorian Yapinus, Alva Erwin, Maulahikmah Galinium, Wahyu Muliady, “Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive Extractive Summarization Technique”, 6th International Conference on Information Technology and Electrical Engineering, ICITEE-2014.

[3] Yutong Wu, Yang Gao, Yuefeng Li, Yue Xu, “Mining Topical Relevant Patterns for Multi-document Summarization”, International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM-2015.

[4] Rachit Arora, Balaram Ravindram, “Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization”, IEEE-2013.

[5] Atif Khan, Naomie Salim, Haleem Farman, “Clustered Genetic Semantic Graph Approach for Multi-document Abstractive Summarization”, IJEDR-2011.

[6] Yogita K. Desai, Prof. Prakash P. Rokade, “Multi Document Summarization: Approaches and Future Scope”, International Journal of Computer Technology and Electronics Engineering (IJCTEE)-2015.