International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey on Multi-Document Summarizer

Prachi M. Joshi[1], Rupali R. Kadam[2]

[1]Department of Computer and engineering, MITCOE, Kothrud, Pune University, Pune, India

[2]Department of Computer Engineering, Pune University, MITCOE, Kothrud, Pune, India

Abstract: ***Natural language processing provides Text Summarization which is the most popular application for information compression. Text summarization is a process of producing a summary by reducing the size of original document and pertaining important information of original document. There is arising a need to provide high quality summary in less time because in present time, the growth of data increases tremendously on World Wide Web or on user's desktops so Multi-Document summarization is the best tool for making summary in less time. This paper presents a survey of existing techniques with the novelties highlighting the need of intelligent Multi-Document summarizer.***

Keywords: Multi-Document Summarization; Clustering Based; Extractive and Abstractive approach; Ranked Based; LDA Based; Natural Language Processing.

1. Introduction

Natural language processing (NLP) is a field of computer science, artificial intelligence and machine learning with the interactions between computers and human language. The use of World Wide Web and many sources like Google, Yahoo! surfing also increases due to this the problem of overloading information also increases. There is huge amount of data available in structured and unstructured form and it is difficult to read all data or information. It is a need to get information within less time. Hence we need a system that automatically retrieves and summarize the documents as per user need in time limit. Document Summarizer is one of the feasible solutions to this problem. Summarizer is a tool which serves a useful and efficient way of getting information. Summarizer is a process to extract the important content from the documents. In general, the summaries are defined in two ways. They are Single Document Summarization and Multiple Document Summarization. The summary which is extracted and created from single document is called as Single Document Summarization whereas Multiple Document Summarization is an automatic process for the extraction and creation of information from multiple text documents.

The main aim of summarization is to create summary which provides minimum redundancy, maximum relevancy and coreferent object of same topic of summary. In simple words, summary should cover all the major aspects of original document without irrelevancy while maintaining association between the sentences of summary. So, Extractive summarization and Abstractive summarization approach is used. Extractive summarization works by selecting existing words, phrases or number of sentences from the original text to form summary. It picks the most relevant sentences or keywords from the documents while it also maintains the low redundancy in the summary. Abstractive summarization method which generates a summary that is closer to what a human might create. Basically this type of summary might contain words not explicitly present in the original document format. It provides abstraction of original document form in fewer words. This survey covers Cluster Based approach, LDA Based approach and Ranking Based approach. The main aim of Multi-document summarization has been also elaborated. The remaining paper is presented as follows. Section II describes related work in the field of multi document summarization using Cluster Based approach, LDA Based approach and Ranking Based approach, Section III presents final conclusion.

2. Related Work

Multi-Document Summarization is an automatic procedure designed to extract and create the information from multiple text documents about the same topic. The multi-document summarization is a very complex task to make a summary. It is a technique where one summary needs to be merged from many documents. There are number of issues in multi document summarization that are different from single document summarization. It requires higher compression. The present implementation includes development of an extractive and abstractive techniques. A 10% summary may be sufficient for one document but if we need it for multiple documents then it is difficult to get a summary from concatenation process. In most if the research, the researcher works on paragraph extraction or sentence extraction because the group of keywords contains a very low amount of information whereas paragraph or sentences can cover the particular concept of document. There are lots of methods which represent multi-document summarization, but in this paper we mainly focus on Cluster based, LDA based approach and Ranking based approach of multi-document summarization.

2.1 Cluster Based Approach

Core of Cluster Based method provides clustering algorithm which is more effective and it depends on centroid of the cluster. Clustering method mainly involves only three task as pre-processing, clustering and summary generation. The following procedure has to be done before providing input to the clustering method by using pre-processing. Basically, pre-processing steps divided into following points-

Tokenization: It breaks the text into separate lexical words that are separated by white space, comma, dash, dot etc. [3]
Stop words removal: Stop words like a, about, all, etc., or other domain dependent words that has to be removed.[3]
Stemming: It removes suffixes like "s", "ing" and so on from documents.[3]

After Pre-processing, clustering method is applied to generate the summary. A paper on data merging by Van Britsom et al. (2013) [1] proposed a technique based on use of NEWSUM Algorithm. It is a type of clustering algorithm where divides a set of document into subsets and then generates a summary of coreferent texts. It contains three phases: topic identification, transformation and summarization by using different clusters. Summarization uses sentence extraction and sentence abstraction. It is splitting the sources by their timestamps. It is divided into two sets as recent articles and non-recent articles. It is based on score of sentence means if information is more accurate then it is added in summarization. It represents higher result for large summarization but general data merging problem arises when unlimited data is available to merge.

This paper is on multi-document summarization using sentence clustering by Virendra Kumar Gupta et al. (2012) [3] states that sentences from single document summaries are clustered and top most sentences from each cluster are used for creating multi-document summary. The model contains the steps as pre-processing, noise removal, tokenization, stop words, stemming, sentence splitting and feature extraction. Feature extraction involves following steps as-

Precision: It is defined as the fraction of retrieved docs that are relevant given as

Relevant = P(relevant | retrieved) [9]

Pn = m/N-n+1

Recall: Fraction of relevant docs that are retrieved given as

Retrieved = P(retrieved | relevant) [9]

Rn= m/n

TFIDF: Formulae [9]

$$\text{TF (term, document)} = \frac{\text{Frequency of term}}{\text{No of Document}}$$

$$\text{Term Frequency} = \frac{n_j}{\sum_k n_k} \text{[10]}$$

IDF *(inverse document frequency):* It calculates whether the word is rare or common in all documents. IDF (term, document) is obtained by dividing total number of Documents by the number of documents containing that term and taking log of that.

$$\text{IDF (term, document)} = \log \frac{\text{Total No of Document}}{\text{No of Doc containing term}}$$

TF-IDF: It is the multiple of the value of TF and IDF for a particular word. The value of TF-IDF increases with the number of occurrences within a doc and with rarity of the term across the corpus.

TFIDF=TF*IDF

After performing these steps, important sentences are extracted from each cluster. And for this, there is two types of sentence clustering used as syntactic similarity and semantic similarity. British National Corpus is used for calculating the frequency of words. It contains 100 million words. It gives best performing system result on DUC 2002 dataset but it is not worked on DUC 2005 or DUC 2006 dataset.

A paper on Extracting Summary from Documents Using K-Mean Clustering Algorithm by Manjula K. S. et al. (2013) [7] proposed K-MEAN algorithm and MMR(Maximal Marginal Relevance) method which are used for query dependent clustering of nodes in text document and finding query dependent summary, depends on the document sentences and tries to apply restriction on the document sentence to get the relevance important sentence score by MMR known as generic summarization approach. Summary of document can be found by k-mean algorithm. This method used to process the dataset by using some clusters and finds prior in the datasets. This helps to find similarity of each document and create the summary of the document. In this work, n-gram which is subtype of co-occurrence relation is used. This helps to process the data set through certain number of clusters and find the prior in the data sets but MMR depends on the document sentences, and tries to apply restriction on the document sentence.

This paper is on Context Sensitive Text Summarization Using K Means Clustering Algorithm by Harshal J. Jain et al. (2012**)** [12] represents K-MEAN algorithm. K-mean clustering is used to group all the similar set of documents together and divide the document into k-cluster where to find k centroids for each cluster. These centroids are not arranged properly so it gives different result. So, we place it properly to group the nearest centroid. Thus we repeat this step until the completion of grouping to the entire document. After this we have to re-calculate k new centroid by considering the center of previous step clusters. These k new centroids generate the new data set point of nearest new centroid. Here loop is generated and k-centroids change their place step by step until any changes are occurred. It finds query dependent summary. Effectiveness and time consumption is the main issues in this approach.

This paper is on Word Sequence Models for Single Text Summarization by Rene Arnulfo Garcia-Hernandez et al. (2009) [13] proposed the Extractive summarization technique which provides a summary to the user for similar text documents. In this paper, here also employs the n-gram(non-grammatical) which consists of sequence of n words within a certain distance in the text and consecutively appear in the text. N-gram is used in a vector space model in determining the extractive text summarization. When sequence of two or three words is used then their probabilities are estimated from a CORPUS which consists of set of documents. At the last, the probabilities are combined to get a priori probability of most probable interpretation. In this work, n-gram is used as a feature of a sentence in an unsupervised learning method. This method is used for clustering the similar sentences and forms the clusters where most representative

sentences are chosen for generating the summary. The algorithm defined as follows-

- *Pre-processing*- First, eliminate stop words, remove noise and then apply stemming process on it.
- *Term selection*- decision has to be taken that which size of n-grams as feature are to be used to represent the sentences. The frequency threshold was 2 for MFS model.
- *Term weighting*- decision has to be taken that how each features are calculated.
- *Sentence clustering*- decide the input for the k-mean algorithm.
- *Sentence selection*- After finishing k-mean algorithm; choose the nearest sentence to each centroid for generating the summary. It provides a summary to the user for similar text documents. It is necessary to find a priori way of determining the best gram size for text summarization what is not clear how to do.

2.2 Ranking Based Approach

Ranking Based Approach usually provides the higher ranked sentences into the summary. Ranking algorithms extracts the rank sentences and merges the all rank sentences and generate the summary. Basically, it applies ranking algorithm, extracts rank sentences and generate a summary.

This paper on SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization by Su Yan and Xiaojun Wan (2014) [19] explain a strategy that it ranks sentences by using SR-Rank algorithm on Extractive text summarization. SR-Rank algorithm is a type of graph based algorithm. Firstly, assign the sentences and get the semantic roles, and then apply a novel SR-Rank algorithm. SR-Rank algorithm simultaneously ranks the sentences and semantic roles; it extracts the most important sentences from a document. A graph based SR-Rank algorithm rank all sentences nodes with the help of other types of nodes in the heterogeneous graph. Here three kinds of graphs are explained as graph-cluster, graph-scan and basic graph. So in this paper, three kinds of graphs are generated as SR-Rank, SR-Rank-span and SR-Rank-cluster. Experimental results are given on two DUC datasets which shows that SR-Rank algorithm surpasses few baselines and semantic role information is validated which is very helpful for multi-document summarization.

Another paper Document Summarization Method based on Heterogeneous Graph by Yang Wei (2012) [20] explains the Ranking algorithm that applies on heterogeneous graph. Existing technique mainly uses statistical and linguistic information to extract the most important sentences from multiple documents where they cannot give the relationship between different granularities (i.e., word, sentence, and topic). The method in this paper actually first applied by constructing a graph which reflect relationship between different granularity nodes which have different size. Then apply ranking algorithm to calculate score of nodes and finally highest score of sentences will be selected in the document for generating summary. By using DUC2001 and DUC 2002, it demonstrates the good experimental result.

A paper on A Novel Relational Learning-to-Rank Approach for Topic-Focused Multi-Document Summarization by Yadong Zhu et al. (2013) [21] provides Optimization algorithm and R-LTR (Learning-to-rank) approach. Relational R-LTR framework is used rather than traditional R-LTR in an elegant way which avoids diversity problem. Diversity is a challenging problem in extractive summarization method. The ranking function specifically define as the combination of ran sentences from documents and for this which is applied first then loss function is applied on Plackett-Luce model which provides ranking procedure on user sentences. Stochastic gradient descent is then utilized to conduct the learning process, and the summary is generated by predicting greedy selection procedure. Quantitative and qualitative approach can be given by experimental results on TAC 2008 AND TAC 2009 which provides state-of-art methods. To accommodate the learning method which will use on other type of dataset beyond the traditional document.

Another paper on Learning to Rank for Query-focused Multi-Document Summarization by Chao Shen, Tao Li (2011) [22] explore how to use ranking SVM to prepare the feature weight for query-focused multi-document summarization. As abstractive summarization provides not well matched sentences from the documents and human generated summary is abstractive so for this reason ranking SVM is applicable here. First, estimate the sentence-to-sentence relationship by considering probability of sentence from the documents. Second, cost sensitive loss function is chosen to make derived training data less sensitive in the ranking SVM's objective function. Experimental result demonstrates effective result of proposed method.

2.3 LDA Based Approach

Latent Dirichlet Allocation (LDA), has been recently introduced for generating corpus topics [22], and applied to sentence based multi-document summarization method. It is not compulsion to estimate topics are of equal importance or relevance collection of sentence or significance themes. Some of the topics can contain different theme and irrelevancy so for this LDA is used for topic model.

The paper Mixture of Topic Model for Multi-document Summarization by Liu Na (2014) [15] based on Titled-LDA algorithm which models title and content of documents then mixes them by asymmetric method. Here mixture weights for topics to be determined. Topic model illustrate an idea how documents can be modelled in the form of probability distributions over words in a document. Titled-LDA divided into three tasks: First, distribution of topic is done over the topic which is sampled from a Dirichlet distribution. Second, a single topic is selected according to this distribution for each word in the document. Finally, each word is sampled from a polynomial distribution over words which are defined in sampled topic. And get the title information and the content information in appropriate way which is helpful in performance of Summarization. The experimental results shows good result by proposing a new algorithm compared to other algorithm on DUC 2002 CORPUS.

The paper Latent Dirichlet Allocation and Singular Value Decomposition based on Multi-Document Summarization by Rachit Arora et al. (2008) [16] proposed LDA-SVD (Latent Dirichlet Allocation and Singular Value Decomposition) Multi-Document Summarization algorithm. As multi-document summarization covers different events from the sentences in the documents and LDA break down that documents into different topics or events. But here orthogonal vector is required to reduce common information content and it provides association of sentences. SVM is used to get the orthogonal representations of vectors and also can represents in the form of sentence orthogonal. LDA finds different topics in the documents whereas SVD finds the sentences which are best represent these topics. Finally, evaluate the algorithms on DUC 2002 CORPUS multi-document summarization tasks using the ROUGE evaluator to evaluate the summaries. This algorithm gives better results for ROUGE-1 recall measures in comparison of DUC 2002.In this. LDA-SVD Multi-Document summarization algorithm is better than GISTEXTER and WSRSE.

This paper Multi-document Summarization based on Hierarchical Topic Model by Hongyan Lill et al. (2011) [17] represents h-LDA (hierarchical Latent Dirichlet Allocation) algorithm introduced for extractive multi-document summarization method. h-LDA algorithm divide into four steps as Pre-processing of the data set, Sentence weighting, Similarity Calculation and Summary sentence compression. It represents productive probabilistic model. This extracts latent topics from multiple documents and also can organize these topics into a hierarchy to gain semantic analysis. At the same time sentence compression technology is used to precise the summaries. So by doing this, we get concise summary. Here TAC 2010 datasets are used for experimental purpose and also ROUGE method is used for evaluating the results. It gives better results than traditional method.

The paper on Topic-Sensitive Multi-document Summarization Algorithm Liu Na et al. (2014) [18] proposes Topic-Sensitive Multi-Document Summarization algorithm. This algorithm divides the topic into two categories as significant topic and insignificant topic. Significant topic as LDA character of sentence method is used in this proposed model for checking similarity between sentence topic. This approach highlights the benefits of statistics characteristics and cooperated with LDA topic model. LDA feature is used to calculate sentence weight. This approach provides better result using DUC 2002 CORPUS as compared to other state-of-art algorithms.

3. Conclusion

In this paper, concepts of multi-document summarization are reviewed with different approaches. This literature review examines the recent trend in summarization system and natural language processing is used to create the summary which is based on human interaction and computer system. Almost all techniques used in summarization that provides correlated information about the topic. There is association found after summarization of multiple documents. Around 22 papers have been discussed here and other techniques that is already exists that also described in this survey. From over all survey, it is clear that multi document summarization is better technique than single document summarization. So, anyone can get a new direction for better perception which will help to construct a new procedure for next age.

References

[1] Van Britsom, Daan, Antoon Bronselaer, and Guy De Tre. "Using data merging techniques for generating multi-document summarizations." *in IEEE trans. On fuzzy systems, pp 1 -17, 2013.*

[2] Bagalkotkar, A., Kandelwal, A., Pandey, S., &Kamath, S. S. (2013, August). A Novel Technique for Efficient Text Document Summarization as a Service.In*Advances in Computing and Communications (ICACC), 2013 Third International Conference on* (pp. 50-53). IEEE.

[3] Gupta, V. K., &Siddiqui, T. J. (2012, December). Multi-document summarization using sentence clustering.In *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on* (pp. 1-5).IEEE.

[4] Ferreira, Rafael, Luciano de Souza Cabral, Rafael DueireLins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J. Simske, and Luciano Favaro. "Assessing sentence scoring techniques for extractive text summarization." *Expert systems with applications* 40, no. 14 (2013): 5755-5764.

[5] Guran, A., N. G. Bayazit, and E. Bekar. "Automatic summarization of Turkish documents using non-negative matrix factorization." In *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on*, pp. 480-484.IEEE, 2011.

[6] ShashiShekhar "A WEBIR Crawling Framework for Retrieving Highly Relevant Web Documents: Evaluation Based on Rank Aggregation and Result Merging Algorithms" in Conf. on Computational Intelligence and Communication Systems, pp 83-88 ,2011.

[7] Manjula.K.S *"Extracting Summary from Documents Using K-Mean Clustering Algorithm"* in IEEE IJARCCE, pp 3242-3246, 2013.

[8] Gawali, Madhuri, MrunalBewoor, and SuhasPatil. "Review: Evaluating and Analyzer to Developing Optimized Text Summary Algorithm."

[9] P.Sukumar, K.S.Gayathri "Semantic based Sentence Ordering Approach for Multi-Document Summarization" in IEEE IJRTE, pp 71-76, 2014.

[10] JinqiangBian "Research On Multi-document Summarization Based On LDA Topic Model" *in IEEE Conf. On Conference on Intelligent Human-Machine Systems and Cybernetics* ,pp 113-116 , 2014

[11] Li, Yuhua, David McLean, Zuhair A. Bandar, James D. O'shea, and Keeley Crockett. "Sentence similarity based on semantic nets and corpus statistics."*Knowledge and Data Engineering, IEEE Transactions on* 18, no. 8 (2006): 1138-1150.

[12] Harshad Jain et. al. "Context Sensitive Text Summarization Using K Means Clustering Algorithm" IJSCE, pp no 301-304, 2012.

[13] García-Hernández, René Arnulfo, and YuliaLedeneva. "Word Sequence Models for Single Text Summarization."In *Advances in Computer-Human*

Interactions, 2009.ACHI'09. Second International Conferences on, pp. 44-48. IEEE, 2009.

[14] Ferreira, R., de Souza Cabral, L., Lins, R. D., Pereira e Silva, G., Freitas, F., Cavalcanti, G. D., ...&Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, *40*(14), 5755-5764.

[15] Liu Na et al."Mixture of Topic Model for Multi-document Summarization" In *2014 26th Chinese Control and Decision Conference (CCDC)*, IEEE, pp no 5168-5172.

[16] RachitArora et al. "Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization" *In2008 Eighth IEEE International Conference on Data Mining*, pp no 713-718.

[17] HongyanLill et al. "Multi-document Summarization based on Hierarchical Topic Model" HongyanLill, pp no 88-91.

[18] Liu, N., Tang, X. J., Lu, Y., Li, M. X., Wang, H. W., & Xiao, P. (2014, July). Topic-Sensitive Multi-document Summarization Algorithm. In *Parallel Architectures, Algorithms and Programming (PAAP), 2014 Sixth International Symposium on* (pp. 69-74). IEEE.

[19] Yan, Su, and Xiaojun Wan. "SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization." *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* 22, no. 12 (2014): 2048-2058.

[20] Yang Wei "Document Summarization Method based on Heterogeneous Graph" *In 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012)*, pp no. 1285-1289, 2012.

[21] Zhu, Yadong, Yanyan Lan, Jiafeng Guo, Pan Du, and Xueqi Cheng. "A Novel Relational Learning-to-Rank Approach for Topic-Focused Multi-Document Summarization." In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pp. 927-936. IEEE, 2013..

[22] Chao Shen, Tao Li "Learning to Rank for Query-focused Multi-Document Summarization" *In 2011 11th IEEE International Conference on Data Mining*, pp no.626-634.