ISSN 2278-3091
Volume 10, No.6, November - December 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse111062021.pdf
https://doi.org/10.30534/ijatcse/2021/111062021

A Systematic Survey on Multi-document Text Summarization

Asha Raj[1], Sudheep Elayidom M[2], David Peter S[3]
[1]Cochin University of Science and Technology, India, asharaj.j@gmail.com
[2]Cochin University of Science and Technology, India, sudheepelayidom@gmail.com
[3]Cochin University of Science and Technology, India, davidpeter123@gmail.com

Received Date : October 10, 2021 Accepted Date : November 11, 2021 Published Date : December 06, 2021

ABSTRACT

Automatic text summarization is a technique of generating short and accurate summary of a longer text document. Text summarization can be classified based on the number of input documents (single document and multi-document summarization) and based on the characteristics of the summary generated (extractive and abstractive summarization). Multi-document summarization is an automatic process of creating relevant, informative and concise summary from a cluster of related documents. This paper does a detailed survey on the existing literature on the various approaches for text summarization. Few of the most popular approaches such as graph based, cluster based and deep learning-based summarization techniques are discussed here along with the evaluation metrics, which can provide an insight to the future researchers.

Key words: Single document, multi-document, cluster based, graph based, deep learning-based text summarization.

1. INTRODUCTION

Text summarization has become a crucial and timely tool for comprehending text content. As Internet frequently delivers more and more information, manually summarizing these massive amounts of text is quite challenging for humans. The main purpose of automatic text summarization is to reduce the length of the source text while preserving the information content and the overall meaning [27]. This can also save the reading time. A good summary should reflect the diverse topics of the document while maintaining the redundancy to a minimum. Multi-document text summarization [30] generates a summary from multiple documents, each of which covers a different perspective and was created at different times. It is more challenging to perform multi-document text summarization because there is more diverse and conflicting information among the documents. The relationships between the documents are also more complicated. Some of the real-world applications of multi-document summarization [21] includes news summarization, scientific paper summarization [16], product reviews, article generation, etc.

Text summarization can be classified based on different categories [23] as shown in Figure 1. Based on the number of documents involved, it can be classified into single document summarization [10] and multi-document summarization [11]. Based on the characteristics of summary generated, it can be classified into extractive based summarization [14] and abstractive based summarization [24]. In single document summarization, the summary is generated from a single document, whereas multi-document summarization [13] takes in a group of documents to generate the summary. Extractive text summarization generates summary, without changing the original text, by extracting proper set of sentences from a single document or multiple documents. Abstractive text summarization generates summary by using new phrases and sentences to capture the meaning of the source document.

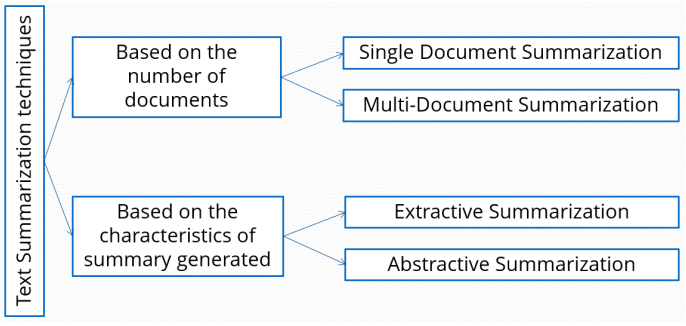

Figure 1: Text Summarization Techniques

This survey paper is organized as follows: Section 2 reviews the graph-based techniques of single document text summarization. Section 3 has a brief review of cluster-based text summarization models in single as well as multiple documents. Section 4 reviews the deep learning-based techniques for text summarization. Section 5 provides a

comparative study of the techniques discussed here and finally section 6 concludes the survey.

2. GRAPH BASED METHODS

Graph-based methods are completely unsupervised method in which a graph is constructed consisting of vertices and edges [22]. In case of single document summarization, sentences are represented as vertices; whereas in multi-document summarization, each document is represented as vertices. If two vertices are related to each other (share common information), they are connected using edges. Edges can be weighted or unweighted, and graphs can be directed or undirected. Graph-based approaches [20] rely solely on the text to be summarized and require no training data. Google's Page Rank algorithm and Kleinberg's HITS algorithm are two of the most prominent graph-based ranking algorithms [19].

2.1 TEXTRANK APPROACH

TextRank is a graph based ranking model for text processing. The graph-based approach proposed by R.Michalcea et.al [1] for text summarization uses this approach. The authors have introduced two unsupervised methods for keyword extraction and sentence extraction. Graph based ranking algorithms are mainly used to determine the importance of a vertex within a graph. Vertices can be either words in a text (for keyword extraction) or it can be an entire sentence (for sentence extraction). This algorithm will assign a score to each of the vertices based on its importance. Scoring of a vertex can be done using a voting system (for keyword extraction) or a recommendation system (for sentence extraction) [12]. In case of a directed graph, the score for a vertex is calculated based on the number of incoming edges and the number of outgoing edges of that vertex. In an undirected graph, if two vertices are connected by an edge, each vertex is casting a vote for the other vertex. The importance of a vertex is determined based on the number of votes casted for that vertex. The higher the number, the more important the vertex (keyword/sentence) is. For keyword extraction, the dataset used is the 'Inspec database' and the results are evaluated using precision (31.2), recall (43.1) and f-measure (36.2). For sentence extraction, the summary generated is evaluated using the ROUGE evaluation toolkit (0.4708). This approach is capable of generating both short and long summaries.

2.2 EXTRACTION BASED APPROACH USING SHORTEST PATH ALGORITHM

S. Jonas et.al [2] proposed an extraction-based summarization technique using shortest path algorithm. The text to be summarized is first split into sentences and words. A graph is then constructed with sentences as nodes. Two sentences are connected using an edge if they are similar to each other. Similarity is calculated by finding the word overlap between two sentences. That is, there will be an edge between two sentences if they share at least one word. Weights are also assigned to the edges based on the similarity. The more similar two sentences are, the less the cost of that edge will be. This method will work only if there is a path from the first sentence to the last sentence. Among the several paths, the shortest path which is closer to the desired length is selected for summary generation. This algorithm ensures a better flow in the summary generated which can't be ensured by any other extraction-based approaches. The dataset used is DUC2005 and the results are evaluated using ROUGE (17) automatic evaluation method.

3. CLUSTER BASED METHODS

A cluster of documents can be considered as a network of sentences that are related to each other [22]. Some sentences are more similar to each other while some others may share only a little information with the rest of the sentences. The sentences that are similar to many of the other sentences in a cluster are more central (or salient) to the topic. Clustering based summarization [28] uses some of the similarity measures like cosine similarity, sentence similarity, Jaccard similarity, support vector machine, etc.

3.1 DOCUMENT AND SENTENCE CLUSTERING APPROACH

A.R. Deshpande et.al [3] proposed a new query based clustering approach for multi-document text summarization. Documents to be summarized are extracted based on the user query. It first groups documents into different document clusters based on cosine similarity. Within each document cluster, similarity of sentences is calculated based on which the sentences are grouped into different sentence clusters. This method generates a cluster of clusters. This ensures greater coverage of the topic and at the same time reduces redundancy. The similarity between sentences is calculated based on cosine similarity measure. Features such as length and position of the sentence, Term Frequency and Inverse Document Frequency, similarity with other sentences, noun feature, etc are used for sentence scoring. The score for each of the sentence cluster is calculated. Sentences with highest scores from each group (sentence cluster and document cluster) are selected for summary generation. The proposed method is evaluated using precision (0.57), recall (0.47) and f-measure (0.52).

3.2 SELF-ORGANIZING MAP CLUSTERING

M.R. Rahul et.al [4] proposed a new technique for extractive single document summarization in Malayalam which uses the concept of semantic role labelling and Self Organizing Maps (SOM). Initially, entity recognition is done where the words

in the document are categorized into 3 main categories: entity, sub-entity and non-entity words. In order to extract the most important sentences from the input document, relevance analysis is done. Each sentence will be assigned a score by considering the following sentence scoring features: sentence entity score, frequent pattern score and semantic similarity score. The scored sentences are then clustered using self-organizing maps where the redundant sentences will be group together. From these clusters, relevant sentences are extracted. Datasets used are some collections of articles from websites such as manoramaonline.com and Wikipedia. The proposed method is evaluated using metrices like precision (0.8), recall (0.667) and F-measure (0.738) against some of the online summarizers.

3.3 Graph And Cluster Based Hybrid Approach

J. Zhao et al. [5] proposed an unsupervised method for multi-document text summarization. Initially, as part of text pre-processing, all the documents are merged and the sentences are extracted. By considering the lexical and deep semantic relationships between these sentences, a sentence graph is constructed with nodes representing the sentences generated and edges are drawn based on 4 main concepts: deverbal noun reference, entity continuation, discourse markers and sentence similarity. Spectral clustering method is applied on the sentence graph to generate multiple clusters of sentences. A summary is generated from each of these clusters and a multi-sentence compression method is applied to generate the final summary. The proposed algorithm is evaluated using ROUGE scores on multi-news (R1 score:42.32) and DUC2004 datasets (36.30).

4. DEEP LEARNING BASED METHODS

Deep learning models are capable of solving complex non-linear relationships [26]. It has been widely used in many domains such as computer vision and natural language processing. Better performance can be achieved by utilizing deep learning-based approach for multi-document text summarization [15].

4.1 Recurrent Neural Network Model

Xin Zheng et.al [6] proposed a hierarchical Recurrent neural network (RNN) model for extractive subtopic-driven multi-document text summarization. RNN models are best in handling sequential data. It is assumed that the documents to be summarized belongs to the same topic, but can contain different subtopics. These sub topics can be present across several input documents. Sentence salience is calculated by considering both subtopic salience and relative sentence salience. Attention mechanism is used to estimate subtopic salience. Similarly, for each subtopic, relative sentence salience is estimated by using the contextual information. Sentences are ranked by multiplying these two values and top ranked sentences are extracted for summary generation. The model is evaluated on two datasets – RA-MDS and DUC2004 and has achieved a ROUGE score of 0.456 and 0.443 respectively.

4.2 Hybrid Model Using Pointer Generator Network & LexRank Approach

A. K. Singh et.al [7] proposed hybrid architecture for multi-document summarization by cascading both abstractive and extractive approaches. Earlier, abstractive summarization techniques were used only to generate headlines for news articles. Later using deep learning techniques, abstractive summaries were generated for single document. In this proposed method, a hybrid approach is used. First, abstractive summarization using pointer-generator technique is applied on multiple large document which creates multiple short and abstract summaries. Later, extractive summarization using LexRank technique is applied that selects important sentences from these summary documents to generate the final summary. This approach ensures greater coverage of topic and also reduces redundancies. When compared to an extractive multi-document summarization, the proposed framework achieved better ROUGE scores(R1:0.4301) when applied on DUC 2004 dataset containing news articles on various topics.

4.3. Transformer Model

The transformer architecture is proposed by Google in 2017 which makes use of attention layer in the encoder-decoder model. Each of the encoder-decoder layer is connected to an attention layer which helps in remembering the position and sequence of words in the input sequence and assigns a weight to it. Hugging face, pipeline, BART (Bi directional and Auto Regressive Transformers), T5(Text-to-Text Transfer Transformer) and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence models) are different models that are based on Transformers. Anushka Gupta et.al [8] did a comparison study on these pre trained models. They used the BBC news dataset for the analysis and found out that the T5 model (ROUGE score:0.47) outperformed all other models.

4.4. Reinforcement Learning

The main approach of reinforcement learning is to take suitable actions in order to maximize the reward. The learner is not told what actions to perform. Actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. Trial-and-error search and delayed rewards are the two most important

distinguishing features of reinforcement learning. Shashi Narayan et.al [9] proposed an architecture consisting of a sentence encoder, document encoder and a sentence extractor for extractive summarization. It uses reinforcement learning for ranking sentences. Sentence encoder is built using Convolutional Neural Networks (CNN) [29] for identifying salient features in the source document. Document encoder uses Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) [18] for training long sentences. Sentence extractor is also implemented using RNN with LSTM which labels sentences as either 1(important for summary generation) or 0 (not so relevant). The model is evaluated using pyrouge (a python package used to compute all ROUGE values) and has achieved an R1 score of 30.4 on CNN dataset and 41.0 on DailyMail dataset.

5. COMPARATIVE STUDY

A comparative study of various text summarization techniques is discussed in Table 1.

Table 1: Comparative study of various text summarization techniques

Authors[Ref]	Approach	No. of input documents (Single/Multi-document)	Type of summarization (Extractive/ Abstractive)	Classification techniques (Supervised/ Unsupervised)	Dataset used	Evaluation metrics (Precision P, Recall R, f-measure F, ROUGE R1)
Michalcea et al.[1]	Graph based	Single document	Extractive	Unsupervised	Inspec database	P-0.312, R-0.431, F-0.362, R1-47.08
Jonas et al.[2]	Graph based	Single document	Extractive	Unsupervised	DUC2005	R1-35
Deshpande et al.[3]	Clustering based	Multi-document	Extractive	Supervised	News articles	P-0.57, R-0.47, F-0.52
Rahul et al.[4]	Clustering based	Single document	Extractive	Supervised	manoramaonline.com wikipedia	P-0.8, R-0.667, F-0.738
Zhao et al.[5]	Graph and cluster based	Multi-document	Extractive	Unsupervised	Multi news DUC2004	R1-42.32 R1-36.3
Zheng et. Al[6]	Deep learning based	Multi-document	Extractive	Unsupervised	RA-MDS, DUC2004	R1-45.6
Singh et al.[7]	Deep learning based	Multi-document	Hybrid	Supervised	DUC2004	R1-43.01
Gupta et.al[8]	Deep learning based	Single document	Abstractive	Supervised	BBC News Dataset	R1-47
Shashi Narayan et.al [9]	Deep learning based	Single document	Extractive	Reinforcement	CNN DailyMail	R1-30.4 R1- 41.0

It provides a comparison on the various approaches (graph-based, cluster-based or deep learning-based), the number of input documents (single or multi-document), the type of summarization used (extractive or abstractive), the classification techniques (supervised and unsupervised), datasets and evaluation metrics. Most of the approaches used DUC2004 datasets for evaluation. The evaluation measures [25] used were precision, recall, f-measure and ROUGE. Precision is calculated by dividing the total number of sentences that occur in both the candidate summary and reference summary by the number of sentences in the candidate summary. Recall is calculated by dividing the same

total number of sentences that occur in both the candidate summary and reference summary by the number of sentences in the reference summary. F-score measures the harmonic average of precision (P) and recall (R). ROUGE [17] (Recall-Oriented Understudy for Gisting Evaluation) is a package for evaluating the candidate summary with the reference summary. It consists of various metrics such as ROUGE N (ROUGE 1, ROUGE 2), ROUGE L, ROUGE W, ROUGE S and ROUGE SU.

6. CONCLUSION

The importance of text summarization has increased in recent years because of the enormous amount of data available on the internet. Text summarization can be single document or multi-document, abstractive or extractive, depending on the number of documents included and the characteristics of the summary generated. In this paper, various approaches to text summarization such as graph-based, cluster-based and deep learning-based were discussed and analyzed. Deep learning-based approaches are the best for performing abstractive text summarization. This survey also provides insight to future researchers to develop more efficient approach.

REFERENCES

1. R. Mihalcea, P. Tarau. **Textrank: Bringing order into texts**, *in Proceedings of EMNLP*, Vol. 4, Barcelona, Spain, 2004.
2. S. Jonas, A. Kenji. **Extraction based summarization using a shortest path algorithm**, *Conference Proceedings of the 12th Annual Natural Language Processing Conference* NLP2006.
3. A.R. Deshpande & L.M.R.J. Lobo. **Text Summarization using Clustering Technique**, *International Journal of Engineering Trends and Technology (IJETT)*, Volume 4 Issue 8, August 2013.
4. M Rahul Raj, Rosna P Haroon and N V Sobhana. **A novel extractive text summarization system with self-organizing map clustering and entity recognition**, *Sādhanā 45*, 32 (2020). https://doi.org/10.1007/s12046-019-1248-0, 25 January 2020.
5. J. Zhao, M. Liu, L. Gao, Y. Jin, L. Du, H. Zhao, H, Zhang, G. Haffari. **SUMMPIP: Unsupervised Multi-Document Summarization with Sentence Graph Compression**, *in proceedings of 43rd ACM International Conference on Research and Development in Information Retrieval 2020* - Virtual, Online, China, 20 July 2020.
6. Xin Zheng, Aixin Sun, Jing Li and Karthik Muthuswamy. **Subtopic-driven Multi-Document Summarization**, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019 (EMNLP-IJCNLP 2019).
7. Anita Kumari Singh, M Shashi. **Deep Learning Architecture for Multi-Document Summarization as a cascade of Abstractive and Extractive Summarization approaches**, *International Journal of Computer Sciences and Engineering*, Vol.-7, Issue-3, March 2019.
8. Anushka Gupta, Diksha Chugh, Anjum & Rahul Katarya. **Automated News Summarization Using Transformers**, *Sustainable Advanced Computing - Select Proceedings of ICSAC 2021*, April 2021.
9. Shashi Narayan, Shay B Cohen, Mirella Lapata. **Ranking Sentences for Extractive Summarization with Reinforcement Learning**, *Proceedings of NAACL-HLT 2018*, pages 1747–1759, June 2018.
10. Hans Christian, Mikhael Pramodana Agus, Derwin Suhartono. **Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)**, *ComTech Computer Mathematics and Engineering Applications*, Volume.7, Dec 2016.
11. Yong Zhang, Meng Joo Er, Rui Zhao, and Mahardhika Pratama. **Multiview Convolutional Neural Networks for Multidocument Extractive Summarization.**, *IEEE Transactions on Cybernetics 47*, 10, 3230–3242.
12. T. Sri Rama Raju, Bhargav Allarpu. **Text Summarization using Sentence Scoring Method**, *International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056*, Volume: 04 Issue: 04, April 2017.
13. Jorge V. Tohalino, Diego R. Amancio. **Extractive Multi-document Summarization Using Multilayer Networks**, *Preprint submitted to Journal of LATEX Templates*, arXiv:1711.02608v1 [cs.CL] 7 8 November 2017.
14. Rahim Khan, Yurong Qian, Sajid Naeem. **Extractive based Text Summarization Using K-Means and TF-IDF**, *International Journal of Information Engineering and Electronic Business*, May 2019.
15. Logan Lebanoff, Kaiqiang Song, Fei Liu. **Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization**, *arXiv:1808.06218v2 [cs.CL]* 28 Aug 2018.
16. Amanuel Alambo, Cori Lohstroh, Erik Madaus, Swati Padhee, Brandy Foster, Tanvi Banerjee, Krishnaprasad Thirunarayan, Michael Raymer. **Topic-Centric Unsupervised Multi-Document Summarization of Scientific and News Articles**, *Air Force Research Laboratory, USAF, arXiv:2011.08072v1 [cs.CL]* 3 November 2020.
17. Chin-Yew Lin. **ROUGE: A Package for Automatic Evaluation of Summaries**, *Association for Computational Linguistics*, July 2004.
18. Sepp Hochreiter and J¨urgen Schmidhuber. **Long Short-Term Memory**, *Neural Computation 9(8):*1735–1780.

19. Chintan Shah, Anjali Jivani. **Literature Study on Multi-Document Text Summarization Techniques**, *Smart Trends in Information Technology and Computer Communications*, Springer, August 2016.
20. G. Erkan, D.R. Radev. **LexRank: Graph-based Lexical Centrality as Salience in Text Summarization**, *Journal of Artificial Intelligence Research (2004),* 457-479.
21. Abimbola Soriyan, Theresa Omodunbi. **Trends in Multi-document Summarization System Methods**, *International Journal of Computer Applications (0975 – 8887),* Volume 97– No.16, July 2014.
22. Yogesh Kumar Meena, Ashish Jain, Dinesh Gopalani. **Survey on Graph and Cluster Based Approaches in Multi-document Text Summarization**, *IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014),* May 09-11, 2014.
23. Vinit Aghama, Dr.V.K.Shandilyab. **A Survey Paper on Extractive and Abstractive Techniques in Automatic Text Summarization**, *International Journal of Research Publication and Review*s Vol (2) Issue (4) (2021) Page 619-625.
24. PL.Prabha, M.Parvathy. **Extractive and Abstractive Text Summarization Techniques**, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1, May 2020.
25. Josef Steinberger, Karel Jezek. **Evaluation Measures for Text Summarization**, *Computing and Informatics*, Vol. 28, 2009, 251–275.
26. Dima Suleiman and Arafat Awajan. **Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges**, *Research Article*, Volume 2020, Article ID 9365340.
27. Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, Quan Z. Sheng. **Multi-document Summarization via Deep Learning Techniques: A Survey**, *arXiv:2011.04843v2 [cs.CL]* 28 Nov 2020.
28. Xiao-Chen Ma, Gui-Bin Yu, Liang Ma. **Multi-document Summarization Using Clustering Algorithm**, *2009 International Workshop on Intelligent Systems and Applications*, IEEE.
29. Yoon Kim. **Convolutional Neural Networks for Sentence Classification**, *in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, pages 1746–1751, 2014.
30. Yash Asawa, Vignesh Balaji, Ishan Isaac Dey. **Modern Multi-Document Text Summarization Techniques**, *International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878*, Volume-9 Issue-1, May 2020.