Grenze International Journal of Engineering and Technology, Jan Issue

A Systematic Survey of Text Summarization Techniques

Shivangi Kochrekar[1], Neha Kale[2], Darsh Mehta[3] and Sunil Ghane[4]
[1-4]Computer Department, Sardar Patel Institute of TechnologyMumbai, Maharashtra
Email: shivangi.kochrekar@spit.ac.in, neha.kale@spit.ac.in, darsh.mehta@spit.ac.in, sunil ghane@spit.ac.in

***Abstract*—Text Summarization is a technique in which a long, lengthy document can be converted into a brief document while maintaining the crux of information in it. A lot of progress has been made over the past few years in this domain, researchers have been able to perform summarization on single paged as well as multi-page documents. However, the biggest challenge that remains so far is the correct ordering of sentences and then forming a coherent summary from those sentences. In this paper, we endeavor to provide a detailed comparison of the different techniques of summarization put forward by researchers. This paper highlights the different approaches used, along with the results and disadvantages of the same. This will help the researchers to learn more about the topic in a comprehensive manner and identify gaps in approach to work on them and come up with novel approaches to address the issue.**

***Index Terms*— Abstractive Summarization, Extractive Summa-rization, Multi-page Document, Text Summarization.**

I. INTRODUCTION

For today's generation, time is money, people are constantly looking for ways to save time. Nowadays, people want to learn on the go but going through a plethora of information about a particular topic can be abstruse to a novice and even time-consuming. To save both time and effort, [1] researchers are constantly working in the field in of text summarization proposing various novel methods to summarize hundreds of pages of worth information into a few pages so people can get a gist about it without going through such huge documents.

Despite the progress researchers have made over the last decade in this field, there haven't been many papers published about the same. Even the papers [2] [3] [4] [5] that have been published before have not entirely covered the breadth of the topic and most of them are relatively old. We have divided our approaches into 5 main categories, Abstractive Text Sum- marization, Extractive Text Summarization, Clustering Based Text Summarization, using Fuzzy Logic to summarize text and Graph based Text Summarization. We would discuss the methodology used, results obtained, and limitations about the same.

The generic flow of the process is as follows (Fig 1.):

The remaining paper is organized as follows: In section two, we briefly discuss the work done in text summarization. In section three, we meticulously compared the different approaches used for the same. In section 4 and section 5, Conclusion.

Grenze ID: 01.GIJET.9.1.502

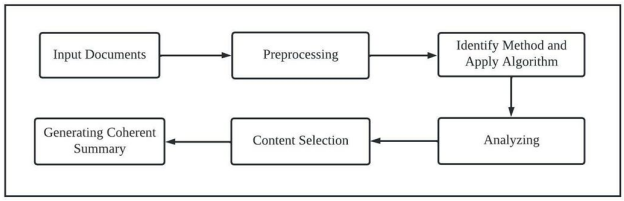

Fig. 1. Generic steps followed for Text Summarization

II. Related Work

Sefid and Giles [6] propose a BERT-based document encoder. Their encoder model stacks multiple transformer layers on sentence vectors to model the inter-sentence relations in the document. This will help build sentence representations. Their SciBERTSUM model, an extension of BERTSUM can generate sentence embeddings for sentences in a document containing multiple sections. It applies a linear sparse attention mechanism between sentences to represent inter-sentence relations.

Ernst et. al. [7] extended clustering-based summarization to apply at the more fine-grained propositional level. First, they extracted all propositions from the input set of documents and filtered out non-salient propositions with a salience-detection model. Then, all salient propositions are clustered into groups based on their semantic similarity. The largest clusters, i.e, those containing information that is more redundant across the documents, are selected to participate in the summary. Finally, each cluster is fused to form a sentence for the abstractive summary.

Du and Huo [8] introduce a novel model for news sum- marization based on multi-feature, fuzzy logic, and the genetic algorithm implemented on Document Understanding Confer- ence for 2002. In the first step, the importance of each word is determined by using features such as word property, word frequency, capital letter, title word, and position of the sentence in which the word is. The weight of each word is calculated. In the next step, time, person, and site characteristics are extracted as authors have assumed they are required while forming a summary. Then the weight of each sentence is calculated based on the number of keywords, the position of the sentence, and the similarity of the sentence with the title. Then the Genetic Algorithm is used to calculate weights and those weights are multiplied by the weight of the sentence calculated to calculate the score of every sentence. The scores are given as input to the fuzzy system which uses the Mamdani Fuzzy Inference method with a trapezoidal membership function. 150 if-then rules were written. After defuzzification, n sentences with the highest scores were selected.

Sotudeh et. al. [9] produce long summaries of scientific papers using three main approaches: Experimenting with var- ious versions of pretrained transformer encoders fine-tuned for the summarization task as an extractive method, a novel multitasking learning model which jointly incorporates the dis- course structure of documents into the extractive summarizer and a two-stage method where an abstractive summarizer is attached to the extractive summarizer to generate abstractive summaries.

Zhang et. al. [10] propose a transformer sentence encoder for document encoding called HIBERT which is short for hierarchical Bidirectional Encoder Representations from Transformers and a technique to pretrain it using unlabeled data. They first use unlabeled data to train the hierarchical encoder of the extractive model. Then the model initialized from the pre-trained encoder learns to classify sentences.

Madhuri and Kumar [11] propose a statistical approach to execute extractive text summarization on a document. Input is given in the form of a text file. Preprocessing of the file is done by firstly tokenizing the files and then stop words are removed from the document. After that, each token is assigned a figure of speech like a noun, pronoun, adjective, and so on. Then each token is assigned a weight based on the number of times it occurs in a document. Individual terms are ranked based on the weighted frequency, which is calculated as a fraction of an individual term's frequency by the term's frequency with maximum frequency. The summary of the document is formed by using the sentences having the highest weight frequency.

Erera et al. [12] describe a novel technique for delivering summaries for Computer Science papers. They determined the most valuable scenarios for scientific document discovery, exploration and developed a model that retrieves and sum marizes scientific documents to satisfy a given information need, either via a query or by selecting categorized values such as scientific datasets, tasks, and more. Their system ingested 270,000 documents. The summary module is designed to produce brief but informative summaries. The summary length of each section was set at 10 phrases for the section-based summary. The section-agnostic summary was

determined tobe of the same length as the section-based summary. 24 papersand 48 summaries were examined in total.

Shirwandkar and Kulkarni [13] propose a unique approach to summarizing text documents by using a combination of Fuzzy Logic and Restricted Boltzmann Machine (RBM). A textdocument is given as input to the system. Then sentences are tokenized and punctuation marks and stop words are taken out.The preprocessed text is used to determine the sentence score based on several sentence features such as the lengthof the sentence, the position of the sentence in the document, number of numerical tokens in the sentence, bigram, and trigram score, number of proper nouns, and keywords in the sentence. The sentence score matrix is passed through theRBM neural network to produce the first summary of the text document. To generate the second summary, the feature scoresare converted into percentages. Then the triangular member-ship function is used for fuzzification, and if-then rules are used for defuzzification which determines the importance of the sentences. For combining both summaries, uncommon and common sentences are found. The final summary includedthe common sentences and half of the uncommon sentences which are sorted according to the position in the document.

Chen et. al. [14] propose the essence vector (EV) model which is a novel unsupervised paragraph embedding method.It aims not only to distill the paragraph's most representative information but also to exclude the general background infor- mation. This generates a low-dimensional, more informative vector representation of the paragraph of interest. Secondly, keeping in mind the increasing relevance of spoken content processing, the denoising essence vector model (D-EV) whichis an extension of the essence vector model is proposed. The D-EV model inherits the benefits of the EV model and infers a more robust rendition of a given spoken passage against flawed speech recognition. Finally, a novel summarization system, that considers both the relevance and the redundancy information simultaneously, is also put forward.

Sethi et. al. [15] describe a method for constructing a summary based on a lexical chain-based model of text topic progression. The basic lexical model is based on the Siber andMcCoy approach because it has linear run time complexity. In the first step, the best representative noun is found for the pronouns used. Then the paragraph is tokenized into sentences and then tokenized into words. A word list is obtained, and pronouns are replaced by the representative noun. The nouns separated are used for lexical chain construction. The chains contain synonymous words according to the WordNet and based on scores, lexical chains having scores greater than the threshold are selected. From the selected lexical chains, individual sentences are scored and sentences having scores above the threshold are selected. Weightage is also given tothe frequency of proper nouns and the first instance of a propernoun having a frequency above a certain value is selected. Thesentences are ordered in the order in which they appear, and the final summary is generated.

Yasunaga et al. [16] propose a neural multi-documentsummarization (MDS) system with sentence relation graphs.A Graph Convolutional Network (GCN) with sentence em-beddings obtained from Recurrent Neural Networks as inputnode features is used on the relation graphs. Using multiplelayer-wise propagation, the GCN generates high-level hiddensentence features for salience estimation. Then, while avoidingredundancy, a greedy heuristic to extract important sentences is employed. Three types of sentence relation graphs are considered in their DUC 2004 experiments. Furthermore, the advantage of combining sentence relations in graphs with therepresentation power of deep neural networks is demonstrated.Ramanujam and Kaliappan [17] propose a new method for multidocument text summarising that combines a timestamptechnique with a Naive Bayesian Classification approach. The timestamp gives the summary an orderly appearance, resulting in a summary that appears to be coherent. It extracts the mostrelevant data from abundant documents. A scoring strategyis used to calculate the score for each word tocalculate the word frequency. In terms of readability and comprehensibility, the higher the linguistic quality, the better. In order to demonstrate the efficacy of the proposed method, this paper compares it to the existing MEAD algorithm. Theproposed method is used to investigate the outcomes of the MEAD algorithm's timestamp procedure.

Saleh et. al. [18] propose a genetic algorithm-based extrac- tive multi-document text summarization model (GA). First,it is treated as a discrete optimization problem, and a specific fitness function is created to give better results with the model. Then, to characterize the adopted GA, representation in binary form is proposed, along with a heuristic mutation and a local repair operator. Experiments are conducted on ten topics from the DUC2002 datasets (Document Understanding Conference)(d061j through d070f).

III. Discussion

The document encoder based on BERT proposed by Sefid and Giles [6] outperforms BERTSUM on the dataset.

The method proposed by Ernst et. al. [7] outperforms the most advanced and modern baselines (in this setting) in both manuals as well as auto evaluation on the DUC and TAC MDSbenchmarks.
When compared with other methods such as Msword, SOM,GCD, Ranking SVM, the algorithm proposed by Du and Huo
[8] gave better results on Rouge-1 and Rouge-2. The drawbackof it is that when a Genetic Algorithm is used it can give local optimum for weights instead of global optimum and for fuzzy logic writing 150+ rules manually is an arduous task.
Sotudeh et. al. [9] found that the new multi-tasking ap- proach outdid the two other methods by huge margins. They report competitive results for their model, which achieves RG(ROUGE) scores of 53.11%, 16.17%, and 20.34% for RG-1, RG-2, and RG-L, respectively.
Zhang et. al. [10] applied the already trained HIBERT to their model and found that it performs better than its randomlyinitialized counterpart by 1.25 ROUGE on the CNN/Dailymaildataset and by 2.0 ROUGE on a version of the New York Times dataset. They also achieved one of the best performances on these two datasets.
The summary generated through the approach of Madhuri and Kumar [11] is compared with a summary generated from MS Word with the proposed methodology having an F1 score of 0.62 while MS Word has an F1 score of 0.53. It is just testedon 5 documents with a maximum of 20 sentences. It uses a weighted average of the documents' words because the context is lost in the process.
For the technique devised by Erera et al. [12], quality of section-based summaries was better than section-agnostic summaries with a score of 3.32 with deviation of 0.53.
The methodology proposed by Shirwandkar and Kulkarni
[13] is compared with the RBM method which had an F1 score of 0.79 while the proposed methodology had a score of 0.84.
On two benchmark summarization corpora, the embedding methods (EV and D-EV) and the summarization framework proposed by Chen et al. [14] were evaluated.
In the model made by Sethi et. al. [15], nouns have been considered to have an important role in the lexical chain, but adjectives also need to be considered. There is no mention of the number of documents that it was tested on, and results are also not mentioned.
The model created by Yasunaga et al. [16] outperforms othercutting-edge multidocument summarization systems, includingtraditional extractive graph-based approaches and the GRU sequence model with no graph.
The findings show that the method proposed by Ramanujamand Kaliappan [17] takes less time to complete the summariza-tion process than the existing MEAD algorithm. Furthermore, the proposed method outperforms the current clustering with lexical chaining approach in terms of precision, recall, and F- score. Because the proposed work does not require a knowl- edge base, it can be used to summaries articles from any field. However, there is a tradeoff between domain independence and a knowledge-based summary, which presents the data ina more user-friendly way.
When compared to another state-of-the-art model, the method by Saleh et. al. [18] proves to be more effective.

IV. Comparative Analysis

In this section, we will compare the existing work andanalyse it in terms of some key features which are shown in TABLE 1.

TABLE I. Approaches for Text Summarization

Ref. No.	Year	Methodology	Results	Gaps
6	2022	SciBERTSUM extends BERTSUM by adding a section embedding layer and a sparse attention technique to integrate section information in the sentence vector.	The recall score was 59.714, 21.498 and 43.057 for ROUGE-L, ROUGE-1 and ROUGE-2 respectively, whichwas higher compared to other base models.	Only a small number of random sentences attend to all other sentences globally, all other sentences,on the other hand, focus on a small number of sentences after and be- fore the current sentence.
7	2021	The models identify the salient propositions, then clusters them into paraphrastic clusters, and finally generates a representative sentence for each cluster by fusing its propositions.	In the TAC 2011 and DUC 2004 datasets, it outperforms the prior state-of-the-art MDS technique when it comes to automatic ROUGE scores and human preference.	It does not work well with long documents.

8	2020	A model was built for news summarization based onmulti feature, fuzzy logic, and genetic algorithm.	It was compared with 8 other methods and tested on ROUGE-1 and ROUGE-2 dataset and it obtained the best results in both cases.	Since fuzzy logic is being used, a lot of human intervention was needed to write about 150 plus if then rules.
9	2020	Pretrained transformer encoders for the extractive sum-marization task using BERTSUM, as well as a multi-task learner that adds a section prediction task to the extractive network, as well as an abstractive summarizer (i.e., BART) that runs over the outputs of the extractive summarizer.	The F-scores of their best models were 45.55, 12.99 and 18.29 for ROUGE-1, ROUGE-2 and ROUGE-L respectively.	On an external dataset, the sum-marization model does not producepromising results.
10	2019	A method for pre-training hierarchical bidirectional trans-former encoders at the document level on unlabeled data.	On the CNN/Dailymail dataset, it surpasses its randomly initialized equivalent by 1.25 ROUGE, and by 2.0 ROUGE on a variant of dataset of New York Times.	The architectures of hierarchical document encoders can be im-proved and other objectives to trainhierarchical transformers can be designed.
11	2019	The top papers based on a query are returned from the corpus. All the sentences are pre-processed using NLTK library and then state-of-the-art extractive, unsupervised, query focused summarization algorithm.	For the section-based sum-maries, the average score was 3.32 with standard deviation of 0.53.	Support for additional entities likemethods is missing.
12	2019	A statistical method of ranking sentences by assigning them weights. Highly ranked sentences are extracted soit extracts important sentences.	The system generated analysis is compared with MS Word generated with f1 scores being 0.62 and 0.53 respectively.	It is just tested on 5 documents with a maximum of 20 sentences. It uses weighted average of the wordsin the documents as a result thecontext is lost in the process.
13	2018	The essence vector (EV) model and the denoising essence vector (D-EV) model are used to create a paragraph embedding framework.	The F-scores of the proposed The model was 0.443, 0.338, and 0.404 for ROUGE-L, ROUGE-1, and ROUGE-2 respectively.	The summary is not coherent.
14	2018	Combination of Fuzzy Logic and Restricted Boltzman Machine is used. Summaries from both the methods are combined at the end to give the result.	The methodology is compared with RBM method which had an f1 score of 0.79 while the proposed methodology had ascore of 0.84.	It is assumed that the first and last sentences have the most impor- tance and very short sentences do not have anything important. This method has been tested on 8 documents only and has average recall and f1 score.
15	2017	A Graph Convolutional Network (GCN) that generates high-level hidden sentence features for salience estima-tion using sentence embeddings obtained from RecurrentNeural Networks as input node features using multiple layer-wise propagation. Then a greedy heuristic to extractimportant sentences is employed.	The model results surpass the previous state-of-the-art multidocument summarization systems, outperforming tradi-tional graph-based extractive techniques and the plain GRU sequence model without a graph.	Semantics have not been taken intoconsideration.
16	2017	Use of lexical chain using a model of topic progression. WordNet thesaurus is used to generate chains. Pronoun resolution is implemented.	The summary generated con-sists of sentences identified by lexical chains and by sentences containing proper nouns.	Adjectives also need to be taken into account for lexical chain. There is no mention of number of documents that it was tested on.
18	2016	The problem is treated as a discrete optimization prob-lem, and a specific fitness function is created to deal with the proposed model effectively. The selected GA is then described using a binary-encoded representation, as well as heuristic mutation and local repair operators.	The average ROUGE scores were 0.263 and 0.087 for ROUGE-2 and ROUGE-L re-spectively.	A base optimization model has been used, the optimization model can be enhanced further.

V. Conclusion and Future Scope

This paper discusses various approaches for text summariza-tion, briefly discussing the methodology used in it alongwith the results and limitations of the same. Promising results are being obtained by using newer Machine Learning techniques but the problem of summarizing too large documents and obtaining a coherent summary persists, hopefully, in a few years, we would be able to overcome those problems as well. This paper could be

used as a baseline for coming years and people wanting to develop more robust algorithms should try and rectify the gaps discussed in this paper.

After this, we will try using the modern summarizationtechniques for an auto survey of research papers and implementa pipeline that will reduce the manual intervention required in a literature survey of papers close to zero.

REFERENCES

[1] Kan, M.-Y., & Klavans, J.L.: Using librarian techniques in automatic text summarization for information retrieval. In: Proceedings of the 2nd ACMlIEEE-CS Joint Conference on Digital Libraries, pp. 36– 45. ACM(2002)

[2] Shah, C., & Jivani, A. (2016, August). Literature study on multidoc- ument text summarization techniques. In International Conference on Smart Trends for Information Technology and Computer Communica- tions (pp. 442-451). Springer, Singapore.

[3] Kumar, Y. J., & Salim, N. (2012). Automatic multi document summa- rization approaches.

[4] Sekine, S., &Nobata, C. (2003, May). A survey for multi-document summarization. In Proceedings of the HLT-NAACL 03 on Text summa-rization workshop-Volume 5 (pp. 65-72). Association for ComputationalLinguistics

[5] Meena, Y. K., Jain, A., &Gopalani, D. (2014, May). Survey on graph and cluster-based approaches in multi-document text summarization. In International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014) (pp. 1-5). IEEE.

[6] Sefid, Athar & Giles, C. (2022). SciBERTSUM: Extractive Summariza-tion for Scientific Documents.

[7] Ori Ernst, &Avi Caciularu, &Ori Shapira, &Ramakanth Pasunuru, &Mohit Bansal, &Jacob Goldberger, &Ido Dagan(2021). A Proposition- Level Clustering Approach for Multi-Document Summarization https://doi.org/10.48550/arXiv.2112.08770

[8] Du, Y., & Huo, H. (2020). News Text Summarization Based on Multi-feature and Fuzzy Logic. IEEE Access, 1–1. doi:10.1109/access.2020.3007763

[9] Sotudeh Gharebagh, S., Cohan, A., Goharian, N.: GUIR @ Long- Summ 2020: Learning to generate long summaries from scientific documents. In: Proceedings of the First Workshop on Scholarly Doc- ument Processing. pp. 356–361. Association for Computational Lin- guistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.sdp-1.41, https://www.aclweb. org/anthology/2020.sdp-1.41

[10] Zhang, X., Wei, F., Zhou, M.: Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. arXiv preprint arXiv:1905.06566 (2019)

[11] Madhuri, J. N., & Ganesh Kumar, R. (2019). Extractive Text Summarization Using Sentence Ranking. 2019 International Conference on Data Science and Communication (IconDSC). doi:10.1109/icondsc.2019.8817040

[12] Erera, Shai & Shmueli-Scheuer, Michal & Feigenblat, Guy & Nakash, Ora & Boni, Odellia & Roitman, Haggai & Cohen, Doron & Weiner, Bar & Mass, Yosi & Rivlin, Or & Lev, Guy & Jerbi, Achiya & Herzig, Jonathan & Hou, Yufang & Jochim, Charles & Gleize, Martin & Bonin, Francesca & Konopnicki, David. (2019). A Summarization System for Scientific Documents.

[13] Shirwandkar, N. S., & Kulkarni, S. (2018). Extractive Text Summa- rization Using Deep Learning. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). doi:10.1109/iccubea.2018.8697465

[14] Chen, K.-Y., Liu, S.-H., Chen, B., & Wang, H.-M. (2018). An Informa- tion Distillation Framework for Extractive Summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(1), 161–170. doi:10.1109/taslp.2017.2764545

[15] Sethi, P., Sonawane, S., Khanwalker, S., & Keskar, R. B. (2017). Automatic text summarization of news articles. 2017 Interna- tional Conference on Big Data, IoT and Data Science (BID). doi:10.1109/bid.2017.8336568

[16] Yasunaga, Michihiro & Zhang, Rui & Meelu, Kshitijh & Pareek, Ayush & Srinivasan, Krishnan & Radev, Dragomir. (2017). Graph-based NeuralMulti-Document Summarization. 452-462. 10.18653/v1/K17-1045.

[17] Nedunchelian Ramanujam, Manivannan Kaliappan, "An Automatic Multidocument Text Summarization Approach Based on Na¨ıve Bayesian Classifier Using Timestamp Strategy", The Scientific World Journal, vol. 2016, Article ID 1784827, 10 pages, 2016. https://doi.org/10.1155/2016/1784827

[18] Saleh, Haruna & Kadhim, Nasreen & Attea, Bara'a. (2015). A Genetic Based Optimization Model for Extractive Multi-Document Text Sum- marization. 56. 1489-1498.