A Survey on Text Summarization Using Natural Language Processing

MANASAVEERASHYVA Y N[1], PRATHIBHA B S[2]
[1, 2] *The National Institute of Engineering, Department of Information Science & Engineering, Mysuru, India*

Abstract*— Text Summarization is a Natural Language Processing (NLP) method that extracts and collects data from the source and summarizes it. Text summarization has become a requirement for many applications since manually summarizing vast amounts of information is difficult, especially with the expanding magnitude of data. Financial research, search engine optimization, media monitoring, question-answering bots, and document analysis all benefit from text summarization. This paper extensively addresses several summarizing strategies depending on intent, volume of data, and outcome. Our aimis to evaluate and convey an abstract viewpoint of the present scenario research work for text summarization.

Indexed Terms*— Natural Language Processing, Text Summarization, Abstractive Summary, Extractive Summary.

I. INTRODUCTION

Since the advancement in the utilization of the Internet has expanded, tremendous volumes of data are generated.Most of the generated data is unstructured, so manually extracting meaningful data from it is challenging [1]. Humans have a constrained ability to comprehend and extract useful information from large amounts of data. It takes a long time for them to grasp the essence of the content. As a result, automatic summarization is a well- known way of addressing such challenges [2].

The objective of text summarization is to gather prominent information from the source by filtering and providing a succinct summary [1]. To date, several techniques for text summarization have been developed. Text summarization techniques can be broadly classified into four categories: input, output, content and purpose. There are single and multi-document summary options based on the number of documents. Meanwhile, the extractive and abstractive outcomes are based on the summary results. In contrast, generic and query-based depend on the purpose [3]. On the other hand, it is divided into indicative and informative based on the content.

The internet is abundant with raw text from several sources, and genres are typically unstructured, noisy, and unsuitable for summary processing [4]. Text pre-processing refers to the process of cleaning and standardizing the unstructured data. It is a necessary step before we can begin text summarizing. The five components of text pre-processing are tokenization, lower casing, stop words removal, stemming, and lemmatization.

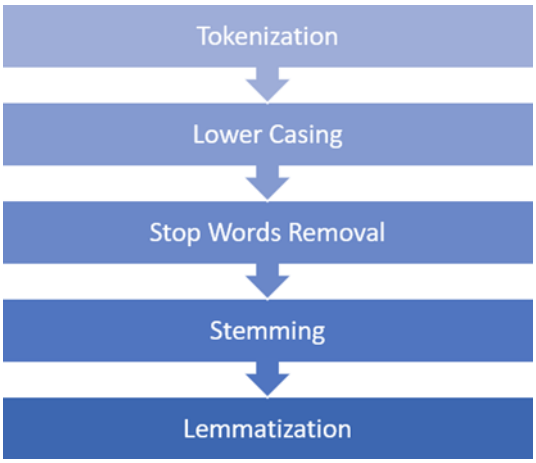

Fig -1: Steps of Text Pre-processing

II. TYPES OF TEXT SUMMARIZATION

i. Extractive vs. Abstractive

Extractive text summarization works by selecting important words, phrase or sentences, and concatenating them to form a meaningful summary. Sentences are chosen based on statistical and linguistic characteristics [5]. Whereas abstractive summarization uses linguistics to examine and interpret the text, and then constructs new sentences

and words while maintaining the source's content in a comprehensible summary [6].

ii. Single Document vs. Multi Document

Single document summarization (SDS) accepts a single document, while multiple document summarization (MDS) accepts several documents as an input. Furthermore, MDS takes into account two categories of documents: homogeneous sets with the same primary context documents and heterogeneous sets with unrelated primary context documents. MDS generates more comprehensive and accurate summaries than SDS, attempting to reconcile different and redundant information [6].

iii. Generic vs. Query-based vs. Domain- specific

Generic-based summaries are independent of the document and may be used by a range of end-users, while query-based summaries are more specific summaries. Domain-specific summaries, on the other hand, leverage knowledge of certain fields, such as scientific and medical publications to develop more comprehensible summaries[7].

iv. Indicative vs. Informative

Indicative summaries contain the metadata of the text. It gives insights of what the document is about and its main idea. While informative summaries provide us with the main background or domain information of the text. It provides information about topic in an elaborated form [4][8]

III. ABSTRACTIVE TEXT SUMMARIZATION TECHNIQUES

Methods	Description	Advantages	Limitation
Word Graph Methodology	The technique based on word graphs is separated into two components. The first component is sentence reduction, followed by sentence combination. This technique involves nodes that represent the information about words and their relation [9].	Word graph technique provides syntactically accurate phrases[10].	The word graph technique creates ungrammatical phrases and is unconcerned with word meaning [10].
Semantic Graph ReductionAlgorithm	The semantic graph-based approach builds a graph that summarizes the original content by gathering semantic information from words and assigning weights to nodes and edges [11].	This method's strength is producing short, coherent, and grammatically accurate phrases with few networks [11].	This approach is restricted to summarizing material from a single document [11].

Markov Clustering Algorithm	To construct summaries, the Markov Clustering Principle employs a hybrid technique. In this method, sentence ranking is accomplished by combining linguistic norms with the best- fitting sentences inside a cluster to construct new sentences [12].	Sentences are grouped using semantic and statistical variables in the Markov Clustering Principle to produce highly linked sentences [12].	The accuracy of the summary provided by the Markov Clustering Principle depends upon the quality of the sentence compression technique [12].

Methods	Description	Advantages	Limitation
Encoder-Decoder Model	The encoder converts the input sentence sequence into a context vector, and the decoder converts the processed input into comprehensible output[13].	The encoder-decoder strength is that it addresses the vanishing gradient issue [14].	The approach requires an extensive dataset that takes a long time to train [11].
Pegasus	In this approach, significant lines are eliminated from the input text and compiled as separate outputs [15].	The strength of this method is that it selects phrases based on relevance rather than randomness [16].	Pegasus may need post-processing to remove errors and enhance summary text output [17].
Summarization with Pointer Generator Networks	This method employs a hybrid approach, producing words from a predefined vocabulary and replicating words by pointing [18].	This method focuses on resolving the issue of out-of-vocabulary terms.	The essence of this technique is to present a summary based on the source content, rather than adding new terminology [19].
Genetic Semantic Graphbased Approach	The approach generates a semantic graph from the source text, with graph nodes representing predicate argument structures (PASs) and graph edges representing semantic similarity weights [20].	The merit of this method is that it reduces redundant information by combining comparable information across documents [9].	The shortcoming of this technique is that it fails to recognize redundant phrases that are semantically similar, resulting in an inadequate final summary [20].

IV. EXTRACTIVE TEXT SUMMARIZATION TECHNIQUES

Methods	Description	Advantages	Limitation

TF-IDF Approach	TF-IDF algorithm calculates the frequency of words in documents and generates metric values. Finally, phrases with a higher metric value are included in the result[13].	The TF-IDF algorithm is quick to compute and has an excellent ability to determine the relevance of phrases [21].	The main disadvantage of TF-IDF is that lengthier sentences get a higher metric score due to the terms' higher occurrence in the sentences [13].
Fuzzy Logic	Fuzzy logic assigns weights to sentences in a document and chooses sentences based on their relevance, determined by sentence length, sentence placement, sentence similarity, and proper noun [27].	The advantage of fuzzy logic is to solve the unequal weighting of attributes to evaluate their relevance [30].	Fuzzy logic cannot solve the issue of dangling anaphora [25].
Approach based on Clustering	The clustering technique focuses on grouping texts and creating cluster-level summaries. The clusters are generated using word weight, sentence location, phrase length, sentence centrality, and proper nouns[22].	The significance of clustering resides in its ability to exclude redundant phrases from summary automatically [23].	The drawback of the clustering approach is that the summarized phrases are not synchronized, and comparing the similarity between clusters is a challenging operation [24].
Neural Network Approach	This method works by first training the neural network, and then the trained network selects the essential phrases that should be included in the summary in the same manner that a person would [5].	The fundamental advantage of neural networks is their ability to change characteristics based on the needs of the user [25].	It takes an excessive amount of time to train a neural network [26].
Approach based on Machine Learning	The machine learning technique is classified into two types: supervised, in which documents and summaries are supplied, and unsupervised, in which just documents are provided, and the machine learns by evaluating them [27].	The benefit of the Machine Learning technique is that it is simple to construct and train the model [28].	The limitation is that significant terms often occur in the test dataset but not in the training dataset are ignored [29].

V. CONCLUSION

Text summarization is a branch of Natural Language Processing (NLP) that focuses on shortening texts and making them more readable for users. With an excess of data accessible on the internet and the necessity to comprehend it in order to save the reader's time, text summary techniques are utilized. This paper provides a quick overview of text preprocessing, used to clean data to do effective summarization. Then it summarizes the many types of textsummarizing approaches, categorizing them according to input,

output, content, and purpose. The paper's primary emphasis is on extractive and abstractive text summarizingalgorithms based on output. Extractive summarization summarizes by simply extracting information from the input text. Abstractive summarization is a more complicated method because it summarizes the text in its language. The abstractive technique produces better and more semantically connected summaries. Readers would benefit significantly from an overview of the benefits and drawbacks of different techniques, as well as a concise explanation. Text summarization techniques can be applied helpfully depending on the user's needs.

REFERENCES

[1] Chen, J., & You, F. (2020, January). Text Summarization Generation Based on Semantic Similarity. In 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 946-949). IEEE.

[2] Dave, H., & Jaswal, S. (2015, September). Multiple text document summarization system using hybrid summarization technique. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 804-808). IEEE.

[3] Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., & Affandy, A. (2020). Review of automatic text summarization techniques & methods. Journal of King Saud University-Computer and Information Sciences.

[4] Rajasekaran, A., & Varalakshmi, R. (2018). Review onautomatic text summarization. Inter. J. Eng. Technol, 7,456-460.

[5] Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. Journal of emerging technologies in web intelligence, 2(3), 258- 268.

[6] Kallimani, J. S. (2018, September). Survey on extractive text summarization methods with multi- document datasets. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2113-2119). IEEE.

[7] Boorugu, R., & Ramesh, G. (2020, July). A survey on NLP based text summarization for summarizing product reviews. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 352-356). IEEE.

[8] Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47(1), 1-66.

[9] Modi, S., & Oza, R. (2018, September). Review on Abstractive Text Summarization Techniques(ATST) for single and multi-documents. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 1173-1176). IEEE.

[10] Talukder, M. A. I., Abujar, S., Masum, A. K. M., Akter, S., & Hossain, S. A. (2020, July). Comparative Study on Abstractive Text Summarization. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE

[11] Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M.,& Kabir, M. M. (2021). A Survey of Automatic Text Summarization: Progress, Processand Challenges. IEEEAccess, 9, 156043-156070

[12] Sahoo, D., Bhoi, A., & Balabantaray, R. C. (2018). Hybrid approach to abstractive summarization. Procediacomputer science, 132, 1228-1237.

[13] Shinde, M., Mhatre, D., & Marwal, G. (2021, March). Techniques and Research in Text Summarization-A Survey. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 260-263). IEEE

[14] Wikipedia Contributors. (2020). Multi-DocumentSummarization— Wikipedia, the Free Encyclopedia. Accessed: Oct. 8, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Multi-document_summarization%&oldid=986613170

[15] Gupta, A., Chugh, D., & Katarya, R. (2021). Automated News Summarization Using Transformers. arXiv preprint arXiv:2108.01064

[16] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020,

November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328- 11339). PMLR.

[17] Oliveira, L. M. R., Busson, A. J. G., Carlos de Salles, S. N., dos Santos, G. N., & Colcher, S. (2021, November). Automatic Generation of Learning Objects Using Text Summarizer Based on Deep Learning Models. In Anais do XXXII Simpósio Brasileiro de Informática na Educação (pp. 728- 736). SBC.

[18] Anh, D. T., & Trang, N. T. T. (2019, December). Abstractive text summarization using pointer-generator networks with pre-trained wordembedding. In Proceedings of the tenth international symposium on information and communicationtechnology (pp. 473-478).

[19] Boutkan, F., Ranzijn, J., Rau, D., & van der Wel, E. (2019). Point-less: More abstractive summarization with pointer-generator networks. arXiv preprint arXiv:1905.01975.

[20] Khan, A., Salim, N., & Kumar, Y. J. (2015, October). Genetic semantic graph approach for multi- document abstractive summarization. In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC) (pp. 173- 181). IEEE.

[21] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268.

[22] Deshpande, A. R., & Lobo, L. M. R. J. (2013). Text summarization using clustering technique. International Journal of Engineering Trends and Technology, 4(8), 3348-3351.

[23] Jewani, K., Damankar, O., Janyani, N., Mhatre, D., & Gangwani, S. (2021, March). A Brief Study on Approaches for Extractive Summarization. In 2021 International Conference on Artificial Intelligence andSmart Systems (ICAIS) (pp. 601-608). IEEE.

[24] Akter, S., Asa, A. S., Uddin, M. P., Hossain, M. D., Roy, S.K., & Afjal, M. I. (2017, February). An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. In 2017 IEEE International Conference on Imaging, Vision & PatternRecognition (icIVPR) (pp. 1-6). IEEE.

[25] Moratanch, N., & Chitrakala, S. (2017, January). A survey on extractive text summarization. In 2017 international conference on computer, communication and signal processing (ICCCSP) (pp. 1-6). IEEE.

[26] Andhale, N., & Bewoor, L. A. (2016, August). An overviewof text summarization technique. In 2016 International Conference on Computing Communication Control and automation (ICCUBEA) (pp. 1-7). IEEE.

[27] Kumar, A. K. S. H. I., & Sharma, A. D. I. T. I. (2019).Systematic literature review of fuzzy logic based text summarization. Iranian journal of fuzzy systems, 16(5),45-59.

[28] Patel, R., Thakkar, A., Makwana, K., & Patel, J. (2017, March). Comprehensive and Evolution Study Focusing on Comparative Analysis of Automatic TextSummarization. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 383-389). Springer, Cham.

[29] Lagrini, S., Redjimi, M., & Azizi, N. (2017). Automatic arabic text summarization approaches. International Journal of Computer Applications,164(5), 31-37.

[30] Babar, M. S. (2014). Improving Text Summarization Using Fuzzy Logic (Doctoral dissertation, RAJARAMBAPU INSTITUTE OF TECHNOLOGY.