A Survey on Text Summarization Techniques

Sneha Thange[1], Jayesh Dange[2], Vivek Karjule[3], Janhavi Sase[4]

Prof. Naina Kokate[5]

Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune.
Savitribai Phule Pune University.

Abstract- In today's digital era, the ever-expanding volume of textual data, abundant on the internet and across various repositories, poses a formidable challenge for manual processing. This paper delves into the fascinating realm of text summarization, a process that distills lengthy content into shorter, more manageable versions. It's like a shortcut for understanding lengthy documents without reading every word. Text summarization refers to creating shorter versions or summaries of lengthy text while maintaining its core idea. This system finds diverse applications, from generating search engine snippets to condensing news headlines, facilitating lawsuit abstraction, and summarizing complex biomedical and clinical texts. Text summarization, particularly for extensive textual documents, presents a significant challenge in the field of natural language processing (NLP). It plays a vital role in NLP by using software to condense lengthy documents into concise summaries that capture the key points. This paper will also introduce different approaches to text summarization, and we will touch on some modern techniques and models that make it all possible.

Index Terms- **Text Summarization, Extractive Summarization, Abstractive Summarization, Text Summarization Models, Performance Metrics.**

1 INTRODUCTION

In today's digital age, the exponential growth of textual data is found abundantly on the internet and in various repositories such as news articles, books, legal documents, and scientific papers. This text keeps growing every day. Reading and summarizing all of this by hand is just too slow and impractical. Text Summarization has emerged as a powerful solution to address this challenge. The Text Summarization system efficiently produces short summaries that capture the core ideas while minimizing redundancy [1]. This system finds diverse applications, from generating search engine snippets to condensing news headlines on websites, facilitating lawsuit abstraction, and summarizing complex biomedical and clinical texts [2]. Text summarization, especially for large textual documents, presents a significant challenge within the field of natural language processing (NLP) [3]. ATS, a vital component of NLP, involves the use of software to condense lengthy text documents into concise summaries that encapsulate the key points [3]. This process can be categorized based on the input type, differentiating between Single Document summarization for shorter texts and Multi Document summarization for longer, more complex inputs [3]. The exponential growth of textual data, both online and in libraries, has made information consumption a time-consuming and challenging task, mainly due to the vast amount of irrelevant content [5]. As a remedy to this issue, Text summarization has emerged as a valuable alternative to manual summarization [5]. Text Summarization aims to provide concise and accurate overviews of lengthy text documents while preserving essential information [5].

Text Summarization methods can be categorized into two primary approaches: Extractive and Abstractive. Extractive summarization involves the selection of important sentences from the input text to compose the summary. In contrast, Abstractive summarization employs advanced natural language techniques to generate its coherent phrases and sentences, closely resembling human-generated summaries [3]. Abstractive methods, while more complex, have made significant progress recently, primarily driven by the utilization of neural networks and models like BERT, GPT, and BART [2].

The remainder of this paper is organized as follows: The literature survey is shown in Section 2, the motivations and applications of text summarization are described in Section 3, text summarization approaches in Section 4, the different models in text summarization are described in Section 5, analysis of recent models in text summarization is described in Section 6, the role of abstractive and extractive approach is explained in Section 7, performance metrics in Section 8. The text summarization challenges with limitations are addressed in Section 9. Finally, Section 10 concludes the paper.

2 LITERATURE SURVEY

In the realm of text summarization, significant contributions have emerged from various studies. Zahoor-ur-Rehman proposed the "Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space" in 2023, which includes combining ROUGE score, TF-IDF, and NLP techniques while emphasizing linguistic feature space [1]. Meanwhile, the Autonomous University of Mexico State's researchers present an innovative extractive summarization method using an encoder-decoder model, as demonstrated in their paper titled "Automatic Generation of an Objective Function for Extractive Text Summarization" (2023), evaluated with ROUGE metrics [2]. Tohida Rehman and Parth Pratimdas, in their work titled "Generation of Highlights from Research Papers using Pointer-Generation Networks and SciBert Embedding" (2023), leverage pointer-generation networks and SciBert embeddings to generate research highlights, showcasing advancements in text summarization techniques [3]. S. Abdur Rahman, P. Mahalakshmi, and N. Sabiyath Fatima delve into the fusion of text summarization and image captioning, employing Bidirectional Long Short-Term Memory (Bi-LSTM) and Deep Belief Networks (DBN) in their paper titled "Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques" (2022) [4]. North China University of Science and Technology researchers enhance the attention-based Bi-LSTM model for hybrid text summarization in their paper titled "Enhancement of Attention-based Bidirectional LSTM for Hybrid Automatic Text Summarization" (2021) [5]. Meanwhile, "Generative Text Summary Based on Enhanced Semantic Attention and Gain Benefit Gate" (2020) by the Civil Aviation University of China aims to elevate summary quality, achieving a ROUGE score of 0.4139 with TF-IDF and an Encoder-Decoder model [6]. Furthermore, Shanghai Normal University's paper, "Abstractive Text Summary Based on Multi-Head Self-Attention" (2019), introduces an innovative abstractive approach integrating multi-head self-attention and CNNs [7]. Lastly, the University of Moratuwa's paper, "Intelligent E-news Summarization" (2018), presents an approach that combines a graph-based and feature-based.

3 MOTIVATION & APPLICATIONS

The study discusses the current state of research in Natural Language Processing (NLP) and Text Summarization. It emphasizes the role of Text Summarization in accelerating knowledge in various fields such as machine learning, natural language, cognitive science, and psychology. Text Summarization enables the creation of new tools, methods, datasets, and resources catering to research and industrial needs. Text Summarization finds applications in summarizing long documents like books, social media posts, sentiment analysis, news articles, emails, legal documents, biomedical documents, and scientific papers [5]. The applications of the Text Summarization domain are listed below:

A. Industry

The use of text summarization in the field of industry is to automatically condense lengthy reports, documents, and data logs, facilitating quicker decision-making and efficient information retrieval.

B. Medical

In the medical domain, text summarization helps summarize extensive medical records, research papers, and patient histories, enabling healthcare providers to access critical information swiftly for diagnosis and treatment decisions.

C. News Media

Text summarization is used in the news media industry to generate concise summaries of news articles and reports, offering readers quick access to the main points of a story.

D. Education

In education, text summarization assists researchers in condensing vast amounts of literature, enabling them to review related works efficiently and stay updated on the latest developments in their field.

E. E-commerce

E-commerce platforms use text summarization to generate product descriptions, reviews, and summaries, providing shoppers with concise information to aid in their purchasing decisions.

F. Legal

In the legal field, text summarization is employed to extract key insights from lengthy court cases, legal documents, and contracts, aiding lawyers and legal professionals in case analysis and research.

4 TEXT SUMMARIZATION APPROACHES

Text Summarization in Natural Language Processing (NLP) is a crucial task that aims to condense and distill the most essential information from a given text while preserving its core meaning. The origins of Text Summarization can be traced back to early information retrieval systems, where the need for concise representations of documents emerged. With the growth of digital content, the importance of automated Text Summarization has gained prominence in various domains.

Text Summarization techniques can be broadly categorized into three main approaches:

a. Extractive approach
b. Abstractive approach
c. Hybrid approach.

A. Extractive Text Summarization:

Extractive Summarization involves selecting sentences or phrases directly from the source text to form a summary. This approach relies on identifying the most informative and representative portions of the original content. Extractive methods often use heuristics, algorithms, or machine learning models to rank and select sentences based on factors such as importance, significance, or redundancy. Early work in extractive summarization includes Luhn's research on automatic indexing and abstracting (1958) [14], which laid the foundation for later developments. There are various types of models such as Graph-based model, Cluster based model, Machine learning Models used in Extractive summarization. These models are elaborated in detail in Section V &VI.

B. Abstractive Text Summarization:

Abstractive Summarization, on the other hand, aims to generate summaries that may contain words or phrases not present in the source text. Abstractive methods require a deeper understanding of the text's content and the ability to generate coherent and concise summaries in natural language. The transition from extractive to abstractive summarization has been driven by advancements in neural network-based models and deep learning techniques [15] [16].

Text Summarization has evolved alongside developments in machine learning and NLP. Earlier approaches relied on rule-based systems and shallow linguistic features [18]. However, recent breakthroughs in deep learning, particularly with the introduction of sequence-to-sequence models and transformers [17], have significantly advanced the state-of-the-art in abstractive summarization. There are various types of models such as sequence-to-sequence model, Pointer Generator Network, Reinforcement Learning Model, and BERT (Bidirectional Encoder Representations from Transformers). These Models are elaborated in detail in Section V and VI.

C. Hybrid Text Summarization:

Hybrid Text Summarization is an approach that combines both extractive and abstractive methods to generate a summary. In a hybrid approach, a system might use extractive methods to identify key sentences or phrases and then employ abstractive techniques to rephrase and consolidate them into a more cohesive summary. Hybrid summarizers tend to produce higher-quality summaries compared to purely extractive methods. This is because they can capture the essence of the text more effectively. Hybrid summarization is less complex than purely abstractive summarization methods. This makes it easier to implement, especially for those who want a balance between simplicity and quality.

5 DIFFERENT MODELS IN TEXT SUMMARIZATION

Text Summarization is a challenging natural language processing (NLP) task, and various models have been used to address it. These models can be broadly categorized into two main types:

a. Extractive Summarization Models
b. Abstractive Summarization Models.

Here's an overview of different models used in Text Summarization:

A. Extractive Summarization Models

1. Graph-Based Models:

- TextRank: TextRank is a graph-based model that treats sentences as nodes in a graph and ranks them based on their importance in the document. Sentences with higher centrality are selected for the summary. It was inspired by Google's PageRank algorithm [19].
- LexRank: LexRank is another graph-based model that uses cosine similarity between sentences to build a sentence similarity matrix. It ranks sentences based on their similarity to other sentences in the document [20].

2. Cluster-Based Models:

Some extractive summarization models group sentences into clusters based on similarity and select one sentence from each cluster to form the summary [21].

3. Machine Learning Models:

Various supervised machine learning models, such as Support Vector Machines (SVMs) and decision trees, have been used for extractive summarization. These models are trained to rank sentences based on features like sentence length, position, and content relevance [22].

B. Abstractive Summarization Models

1. Sequence-to-Sequence Models (Seq2Seq):

Encoder-decoder model is commonly used to implement Seq2Seq tasks. Seq2Seq models consist of an encoder that reads the source text and a decoder that generates the summary. They have been widely used for abstractive summarization tasks. Variants like the Transformer model have achieved state-of-the-art results [23].

2. Pointer-Generator Networks:

Pointer-generator networks combine elements of extraction and abstraction. They can decide whether to copy words from the source text or generate new words to form the summary. These networks were introduced by See et al. in "Get To The Point: Summarization with Pointer-Generator Networks" (2017) [24].

3. Reinforcement Learning Models:

Some abstractive summarization models incorporate reinforcement learning techniques to fine-tune generated summaries. They use reward functions based on metrics like ROUGE to guide the model towards generating better summaries [25].

4. BERT-Based Models:
Pretrained language models like BERT (Bidirectional Encoder Representations from Transformers) have been fine-tuned for abstractive summarization tasks. They generate summaries by attending to the input text and producing coherent and contextually relevant summaries [26].

These are some of the primary models used in Text Summarization. The choice of model depends on the specific task, dataset, and desired output. Researchers continue to explore and develop new models and techniques to improve the quality of Text Summarization.

6 ANALYSIS OF RECENT MODELS IN TEXT SUMMARIZATION

There is a large variety of models available, some of the recent models for effective test summarization are:-

A. BERT

BERT (Bidirectional Encoder Representations from Transformers), in the context of text summarization, is a state-of-the-art pre-trained language model that excels in understanding the nuances of language, making it well-suited for abstractive text summarization. BERT is a part of the abstractive summarization category and is highly effective at producing human-like summaries by generating content that is not explicitly present in the source text. Its effectiveness in text summarization is notable due to its ability to capture context and generate coherent and contextually relevant summaries, though fine-tuning on summarization-specific tasks is often required for optimal performance.

B. Neural Networks

Neural Networks (NN), including various architectures like feedforward networks and convolutional neural networks (CNN), are used in abstractive and extractive summarization tasks. They have proven effective in both categories. In abstractive summarization, NNs are adept at generating summaries by capturing contextual information, while in extractive summarization, CNNs are useful for identifying and extracting salient sentences or phrases. Their effectiveness depends on the specific architecture and training data.

Recurrent Neural Networks (RNNs) are used in text summarization for abstractive tasks, where they can capture the sequential dependencies in the source text. However, RNNs have certain limitations, such as difficulty in handling long-range dependencies and the vanishing gradient problem. In comparison to more advanced models like Transformers, RNNs may not be as effective in generating high-quality abstractive summaries.

C. Graph Based

TextRank and LexRank are graph-based extractive summarization algorithms that rely on the creation of a sentence similarity graph, where sentences are represented as nodes, and edges indicate the similarity between sentences based on certain features. They fall under the category of extractive summarization and are effective for producing concise and coherent summaries while preserving key information in the source text.

D. Bi-LSTM

Bidirectional Long Short-Term Memory networks (Bi-LSTM) are often used in abstractive text summarization. They are capable of capturing bidirectional context in the source text, which is essential for understanding the dependencies between words and phrases. Bi-LSTM models are effective in generating abstractive summaries by considering both past and future information. Their effectiveness is notable in tasks where capturing long-range dependencies is crucial for producing coherent and contextually relevant summaries.

E. PGN

Pointer-generator networks (PGN) represent an advancement in abstractive summarization models. These networks combine elements of both extractive and abstractive approaches. They have the ability to copy words directly from the source text while also generating novel words to form the summary.

F. PEGASUS and T5

Pegasus and T5, developed by Hugging Face, are state-of-the-art pre-trained models that excel in abstractive summarization tasks. These models are highly effective in generating human-like summaries by understanding the context and semantics of the source text. Their effectiveness in abstractive summarization is a testament to the power of transformer-based models, and they have shown superior performance in various summarization benchmarks.

7 ROLE OF ABSTRACTIVE AND EXTRACTIVE APPROACH

The choice between extractive and abstractive summarization for generating research highlights depends on several factors, including the desired level of conciseness, the availability of training data, and the trade-off between accuracy and creativity:

Combination: Some research highlights generation systems use a combination of both extractive and abstractive methods. Extractive methods can provide a factual foundation, while abstractive methods can improve readability and conciseness.

Evaluation: The quality of research highlights generated by both approaches should be evaluated using relevant metrics, ensuring that the generated highlights effectively capture the core contributions of the research paper.

In recent years, advancements in natural language processing, including pre-trained language models like BERT and GPT, have contributed to improvements in both extractive and abstractive summarization, making it possible to generate high-quality research highlights automatically.

Researchers and developers can experiment with various approaches to find the most suitable method for their specific needs.

8 PERFORMANCE METRICS

Performance metrics in the context of text summarization are used to evaluate how well a summarization system or model is performing in generating summaries. These metrics help assess the quality, effectiveness, and characteristics of the generated summaries. Common performance metrics for text summarization include

A. ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

ROUGE-1 F1: Measures the F1 score based on unigram (single word) overlap between the generated summary and reference summaries.

ROUGE-2 F1: Measures the F1 score based on bigram (two consecutive words) overlap between the generated summary and reference summaries.

ROUGE-L F1: Measures the F1 score based on the longest common subsequence between the generated summary and reference summaries.

B. METEOR (Metric for Evaluation of Translation with Explicit ORdering):

METEOR Score: METEOR considers additional linguistic aspects such as stemming, synonyms, and word order. It aligns generated and reference sentences to compute precision, recall, and the harmonic mean (F1) based on unigram matches.

C. BERT Score:

BERT Score F1: BERT Score leverages contextualized embeddings from BERT to measure the similarity between the embeddings of model-generated text and reference text. It calculates precision, recall, and F1 score based on these embeddings.

D. BLUE (Bilingual Evaluation Understudy):

BLUE Score: BLEU is a metric used to evaluate the quality of machine-generated text, including Text Summarization. It measures how closely the generated text matches reference texts. BLEU calculates a score between 0 and 1, with higher scores indicating better quality. It counts matching n-grams (word sequences) between the generated and reference texts, rewarding precision and brevity.

These performance metrics are widely used in the evaluation of Text Summarization systems, and they provide valuable insights into the quality, fluency, and relevance of generated summaries compared to human-written or reference summaries. Researchers often report results using these metrics to assess the effectiveness of different Text Summarization approaches.

9 LIMITATIONS AND CHALLANGES

While Text Summarization has seen significant advancements, it is essential to acknowledge that several formidable challenges continue to confront researchers in this domain. These challenges not only reflect the complexities of the task itself but also point toward promising directions for future studies.

One of the foremost challenges pertains to multi-document summarization. Unlike single-document summarization, this task involves dealing with a multitude of documents, which introduces issues like redundancy, temporal dimension, co-references, and the need for sentence reordering. Furthermore, some approaches in

multi-document summarization can unintentionally generate improper references, contributing to the complexity of this task.

Another set of challenges revolves around the applications of Text Summarization. While many current studies focus on specific text domains, such as news or biomedical documents, some of these domains may lack substantial economic value. This prompts us to consider the economic viability of summarizing longer texts like essays, research papers or reports. However, this approach poses computational challenges due to the high processing power required for handling extensive textual data.

User-specific summarization tasks add another layer of complexity. Summarizing semi-structured resources, such as web pages and databases, holds immense importance as much of the textual data exists in this format. However, the inherent noise in such data makes it more challenging to develop efficient summarization techniques for these domains.

10 CONCLUSION

Text summarization is a vital tool in coping with the ever-expanding volume of textual data in the digital age. It efficiently distills the core ideas from extensive documents, making information consumption more manageable. This paper has explored various approaches to text summarization, including extractive and abstractive methods, and highlighted their applications across diverse fields. Notable models like BERT, NN, and Pegasus have been discussed for their roles in enhancing summarization tasks. The study has emphasized the role of text summarization in accelerating knowledge in various domains, from industry and law to medicine and education. With the continuous development of models and the ever-increasing volume of textual data, text summarization will remain a pivotal component of natural language processing, aiding researchers, professionals, and readers alike.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers of this article and our guide Prof. Naina S. for their valuable suggestions and contribution in helping them to improve the quality of this work.

REFERENCES

[1] Wafaa S. El-Kassasa, Cherif R. Salamaa,b, Ahmed A. Rafeab & Hoda K. Mohameda ,“Automatic text summarization: A comprehensive survey”, 2021, pp. 0957-4174

[2] Ayesha Ayub Syed, Ford Lumban Goal & Tokuro Matsuo, ” A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization”, 2021, pp. 3052783

[3] Ishitva Awasthi , Kuntal Gupta , Prabjot Singh Bhogal , Sahejpreet Singh Anand & Piyush Kumar Soni, “Natural Language Processing (NLP) based Text Summarization - A Survey”, 2021, pp. 978-1-7281-8501-9

[4] Rajat K. Kulshreshtha , Anuranjan Srivastava , Mayank Bhardwaj, ” A Survey Paper on Text Summarization Methods” , 2018, pp. 2395-0072

[5] M. F. Mridha, Kamruddin Nur, Aklima Akter Lima, Sujoy Chandra Das, Mahmud Hasan & Muhammad Mohsin Kabir, “A Survey of Automatic Text Summarization: Progress, Process and Challenges”, 2021, pp. 3129786

[6] Aniqa Dilawari , Muhammad Usman Ghani Khan, Zahoor-Ur-Rehman, Summra Saleem, & Fatema Sabeen Shaikh, “Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space”, 2023 , pp. 3249783

[7] Ángel Hesnandez-Castandea, Yulia Ledeneva & Rene Arnulfo Garcia-Hernandez, “Toward the Automatic Generation of an Objective Function for Extractive Text Summarization”, 2023, pp. 3279101

[8] Tohida Rehman, Plaban Kumar Bhowmick, Debarshi Kumar Sanyal, Samiran Chattopadhyay, & Partha Pratim Das, “Generation of Highlights From Research Papers Using Pointer-Generator Networks and SciBERT Embeddings” , 2023, pp. 3292300

[9] P. Mahalakshmi & N. Sabiyath Fatima, “ Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques” , 2022, pp. 3150414

[10] Zhanlin Ji, Ivan Ganchev, Jiawen Jiang, Haiyang Zhang, Chenxu Dai, Qingjuan Zhao & Hao Feng, “Enhancements of Attention-Based Bidirectional LSTM for Hybrid Automatic Text Summarization”, 2021, pp. 3110143

[11] Jianli Ding, Yang Li, Huiyu Ni & Zhengquan Yang, “Generative Text Summary Based on Enhanced Semantic Attention and Gain-Benefit Gate”, 2020, pp. 2994092

[12] Qian Guo, Jifeng Huang, Naixue Xiong & Pan Wang, “MS-Pointer Network: Abstractive Text Summary Based on Multi-Head Self-Attention”, 2019, pp. 2941964

[13] Varuni Alwis, “Intelligent E-news Summarization”, 2018, pp. 189 - 195

[14] Luhn, H. P. (1958). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 2(4), 309-317.

[15] Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[16] See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL).

[17] Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

[18] Hovy, E., & Lin, C. (1999). Automated Text Summarization in SUMMARIST. Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization.

[19] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[20] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22, 457-479.

[21] Steinberger, J., & Jezek, K. (2004). Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. Proceedings of the European conference on Information Retrieval (ECIR).

[22] Conroy, J. M., Schlesinger, J. D., & O'leary, D. P. (2004). Topic-focused multi-document summarization using an approximate oracle score. Proceedings of the 20th international conference on Computational Linguistics (COLING).

[23] Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

[24] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL).

[25] Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. Proceedings of the 34th International Conference on Machine Learning (ICML).

[26] Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT).