International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Brief Survey on Text Summarization Methods

Joe George Cherian

Department of Computer Science and IT Jain Deemed-to-be University, Bangalore, India j.gc00200@gmail.com

ABSTRACT-

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that enables computers to understand, interpret and generate human language that is both meaningful and relevant to a given context.

Text Summarization is the process of condensing a piece of text, such as an article or a document into a shorter version while keeping the key points and important information.

Text summarization in natural language processing (NLP) is crucial for condensing lengthy texts while retaining vital information. This review paper examines current research and advancements in text summarization techniques. It highlights the significance of summarization across domains like information retrieval and content analysis. The paper explores extractive, abstractive, and hybrid approaches to summarization, along with methodologies for evaluating summaries and commonly used datasets. It also discusses emerging trends such as deep learning models and transformer architectures. Challenges faced by summarization systems are addressed, and future research directions are suggested. With a focus on providing insights into various summarization methods and recent progress, this survey paper aims to be a valuable resource for NLP researchers.

Keywords: *NLP (Natural Language Processing), Text Summarization, transformers, deep learning*

I. Introduction

In today's digital age, the abundance of textual information available on the internet and other digital platforms has led to an increasing demand for effective methods to digest and comprehend large volumes of text efficiently. Text summarization, a fundamental task in natural language processing (NLP), addresses this need by automatically condensing lengthy documents, articles, or text passages into shorter, coherent summaries while retaining the most crucial information. The ability to generate concise summaries from complex textual data has numerous applications across various domains, including information retrieval, document summarization, content analysis, and more.

The primary goal of text summarization is to distill the essential content of a document or text passage, enabling users to quickly grasp the main points without reading the entire document. This capability is especially useful when time is limited, such as browsing news articles, scanning search engine results, or reviewing large volumes of research papers.

Over the years, text summarization has evolved significantly, with researchers developing a wide range of techniques and methodologies to address different summarization tasks and challenges. These techniques can be broadly categorized into two main approaches: extractive and abstractive summarization. Extractive summarization involves selecting and rearranging existing sentences or passages from the original text to create a summary, while abstractive summarization aims to generate summaries by understanding and paraphrasing the content in a more human-like manner.

In this survey paper, I will provide a comprehensive overview of text summarization techniques, evaluation metrics, datasets, and recent advancements in the field. I will explore the strengths and limitations of different summarization approaches, discuss the challenges faced by text summarization systems, and identify future research directions. By synthesizing and analyzing the existing literature on text summarization, this survey aims to provide researchers and practitioners with a valuable resource for understanding the current state of the art and guiding future research efforts.

II. Literature Review

A. Early Approaches

One of the earliest text summarization systems was developed by H.P. Edmundson in the 1960s, which used statistical measures to identify important sentences in scientific articles [1].

With the rise of computers in the 1950s, early attempts at automatic text summarization focused on statistical methods. The weight of the sentences of a given document was a function of the high-frequency words while ignoring common high-frequency words.

A common method in the first statistical techniques involved giving importance to sentences by looking at how often important words appeared there. The important words were found using calculations like term frequency-inverse document frequency (TF-IDF) or counting how many times each word was used. Researchers gave more importance to sentences with these significant words to focus on those most probably holding essential information.

Since then, numerous publications have emerged to tackle the challenge of automated text summarization.

B. Extractive Summarization

Extractive summarization is a technique that involves selecting and extracting important sentences from the original text to create a summary.

C. Abstractive Summarization

Abstractive text summarization is a technique in which a summary of a given text is generated by interpreting and rephrasing the content in a new and concise form, rather than directly selecting and extracting sentences as in extractive summarization.

D. Evaluation Metrics

a. ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE measures how well the summary covers important information from the original text. It looks at the overlap between the summary and the reference summaries (i.e., human-generated summaries). ROUGE calculates several scores, such as ROUGE-N (measuring overlap in n-grams) and ROUGE-L (measuring the longest common subsequence) [2].

b. BLEU (Bilingual Evaluation Understudy): BLEU evaluates the quality of a summary by comparing it to one or more reference summaries. It measures how many n-grams (sequences of words) in the summary match those in the reference summaries. BLEU scores range from 0 to 1, with higher scores indicating better quality summaries [3].

c. METEOR (Metric for Evaluation of Translation with Explicit Ordering): METEOR assesses the overall quality of a summary by considering both content overlap and surface-level similarity. It takes into account word matching, word order, and stemming. METEOR scores range from 0 to 1, with higher scores indicating better quality summaries [4].

E. Datasets

In the study of text summarization, researchers often work with a number of datasets. They use these to teach and test their summary-making models. The data includes different kinds of written material like news stories, research articles, and web content from various fields and styles. Datasets that people often use are made for certain jobs or challenges, like the dataset from Document Understanding Conference and from Text Analysis Conference. Besides these, there are public datasets, for example, the CNN or Daily Mail with news stories and summaries written by humans or PubMed which has many articles on medicine and biology. These datasets give those who research a way to measure how good summarization algorithms are by using consistent standards and they make it possible to compare various methods in an equal manner.

F. Challenges

Summarizing text is hard work with many problems. People who study this are trying very hard to solve them. It's tough to make summaries that tell you enough and that also make sense together. It's often difficult because models that summarize tend to have trouble getting the key points of a text and keeping a sensible sequence in their summaries. Also, working with specialized materials like those for law or medicine is another challenge. Models for summarizing that learn from datasets with many different topics might not be good at summarizing specific content. Also, systems that summarize can find it challenging to work with content coming from more than one document or in different forms where the information is spread out across various places or when different sentences have the same meaning. Another difficulty one may face is dealing with moral issues about prejudice and bias in summarizing. If models are trained using datasets that have biases, these same partialities might show up in the summaries they create. To overcome these difficulties, we need to improve techniques in processing language naturally and consider the impact on society caused by systems that summarize automatically.

III. Research Methodology

A. Goal:

The goal of this paper is to provide an overview of the existing and upcoming techniques in text summarization. At the end of this paper, I will also present the insights that I gained while learning about the field of text summarization.

B. Literature Review Approach:

I plan to carry out a survey of literature by utilizing resources like Google Scholar, IEEE Xplore, and ACM Digital Library. The search will be centered on research papers that describe different methods for summarizing text. For searching, I will use keywords like "text summarization", "automatic summarization", "extractive summarization", "abstractive summarization" and also the phrase “neural text summarization”.

C. Inclusion and Exclusion Criteria

The literature in my survey includes peer-reviewed articles, conference papers, and technical reports that focus on text summarization techniques. Duplicates, non-English publications, and articles that did not focus on text summarization were excluded.

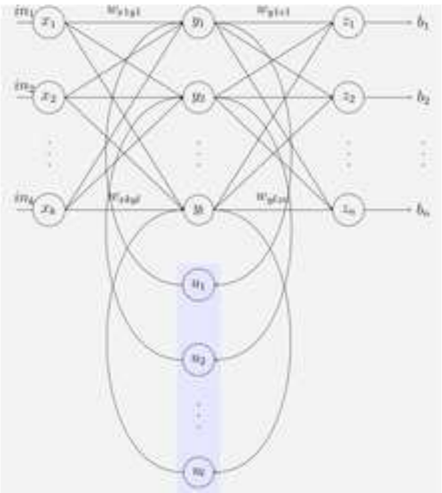
D. Data Analysis

Relevant data was extracted after a thorough evaluation of the chosen literature. This information included datasets (like CNN/Daily Mail, DUC, TAC), evaluation metrics (like ROUGE, BLEU, METEOR), and key findings and summarization techniques (like extractive, abstractive, supervised, unsupervised, etc.). Following the extraction process, the data were categorized thematically according to the kinds of summarization and assessment methods used.

Research Findings

1. Extractive Text Summarization

My findings are that text summarization relies on various different methods, especially when it comes to extractive techniques. These methods cover a wide range of approaches, from using graph algorithms like TextRank and LexRank to more complex machine learning methods such as Support Vector Machines (SVM) and Recurrent Neural Networks (RNN). The real strength of these extractive techniques is their ability to make summaries by figuring out and picking the most important points from the original text. Researchers have come up with all sorts of ways to shrink big chunks of text into short summaries. They use algorithms to study how the text is put together and machine learning models to spot the most important bits. The main goal is to boil down the text into summaries that capture the important points.



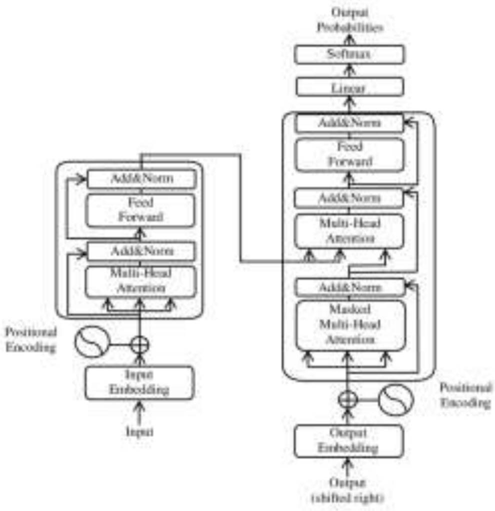
RNN diagram

2. Abstractive Text Summarization

Abstractive summarization techniques aim to generate summaries in their own words. This is generally done using the help of language models that use a transformer architecture (Eg. BERT [5] , GPT [6] or BART [7]).

Language Models like the above are trained on vast amounts of data like books, websites, articles and various other sources. During the training of these models, the model learns to understand the structure, semantics, context and syntax of a language by analyzing the patterns in the text.

BART, BERT and GPT are based on the transformer architecture, which is a type of deep learning model specifically designed for sequence-to-sequence tasks in NLP. The transformer architecture consists of multiple layers of self-attention mechanisms and feed-forward neural networks, allowing the model to capture long-range dependencies and contextual information in text.



Transformer Architecture diagram

The self-attention mechanism is like the powerhouse of the transformer setup. It's what helps the model figure out which words in a sequence are the most important and how they relate to each other. By giving attention to the right bits of the input sequence, the model can direct its attention on the key points and pick up on connections between words, even if they're far apart in the text.

3. Hybrid Text Summarization

Hybrid text summarization aims to combine elements of both the extractive and abstractive text summarization techniques. Hybrid text summarization aims to create a balance between preserving the original words while also improving the readability and coherence. One of the most common approaches to implement hybrid text summarization is by creating an extractive summary first and then using abstractive techniques to rewrite and refine the sentences to create a good quality summary.

Recommendations

Exploring Mamba for Text Summarization

Recent advancements in machine learning models present exciting possibilities for text summarization. One such model, Mamba [8], demonstrates potential advantages over traditional Transformer-based architectures. Mamba could be explored for improved summarization because of its:

- Efficiency and Scalability:

Mamba's faster inference speed and linear scaling with text length could enable real-time summarization of even very long documents. This is particularly valuable for applications requiring summaries of constantly updated content (e.g., news feeds, research articles).

- Improved Attention Mechanisms: Mamba's selective attention mechanism focusing on crucial information aligns well with the core objective of summarization – identifying key points. This focus could lead to more accurate and concise summaries compared to Transformers.
- Integration with Existing Systems: Explore the feasibility of integrating Mamba into existing text summarization pipelines. The potential speedup and potentially higher quality summaries could significantly enhance current summarization capabilities.

A Potential Limitation

- Mamba is a relatively new player compared to established models like Transformers. Although it demonstrates promise, there's a dearth of research and well-defined practices specifically for text summarization with Mamba. This means significant fine-tuning and experimentation might be necessary to optimize Mamba's effectiveness for summarizing text.

Conclusion

In conclusion, text summarization has come a long way from its early beginnings of keyword extraction and statistical methods. The field has seen a surge in advancements with the rise of neural networks, particularly powerful architectures like Transformers. These models have led to significant improvements in capturing the gist of text data and generating concise, informative summaries. As we explored, recent advancements like Mamba hold promise for pushing the boundaries of efficiency and accuracy even further. However, challenges remain, such as effectively handling factual information, mitigating bias, and ensuring summaries remain faithful to the source text's nuance. As research continues to explore new techniques and leverage ever-growing computational power, the future of text summarization is bright, offering exciting possibilities for various applications across diverse fields.

References

[1]“New Methods in Automatic Extracting | Journal of the ACM,” Journal of the ACM, Apr. 01, 1969. https://dl.acm.org/doi/10.1145/321510.321519

[2] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in Text Summarization Branches Out, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[4] A. Lavie and A. Agarwal, “Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments,” in Proceedings of the Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.

[7] M. Lewis et al., BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019.

[8] A. Gu and T. Dao, *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. 2023.