ITM Web of Conferences **73**, 02008 (2025) https://doi.org/10.1051/itmconf/20257302008
IWADI 2024

From Extractive to Generative: An Analysis of Automatic Text Summarization Techniques

Zixu Liu*

School of Artificial Intelligence, Tiangong University, 300000, Tianjin, China

Abstract. With the explosive growth of digital content, the demand for effective information retrieval and summarization has become increasingly important. This paper provides a comprehensive review of automated text summarization techniques, focusing on the challenge of condensing large volumes of text into concise summaries. The article explores the evolution of automatic summarization methods, from early extractive techniques to modern generative approaches based on deep learning. The review highlights significant milestones in the development of summarization algorithms, including the emergence of Transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), which have significantly improved the quality and coherence of generated summaries. Additionally, the paper examines the diverse applications of summarization technologies across fields such as healthcare, discussing the challenges and solutions presented in the literature. By shedding light on the current advancements and ongoing challenges, this review underscores the crucial role of automated text summarization in enhancing information accessibility, with promising implications for future research and practical applications.

1 Introduction

In today's information-rich environment, the ability to quickly process and comprehend vast quantities of textual data has become increasingly crucial. Automatic text summarization technology has emerged as an essential tool for addressing the challenges posed by information overload. By extracting or generating concise representations of larger texts, this technology enables users to swiftly access key insights and relevant information. The rapid expansion of the internet and the exponential growth of digital content have exacerbated the issue of information saturation, making effective filtering and comprehension vital. As a result, automatic text summarization has gained significant attention as a means to enhance information retrieval and user experience.

Over the past few years, advancements in natural language processing (NLP) and deep learning have propelled the evolution of summarization techniques. Initially dominated by rule-based and statistical approaches, the field has seen a shift towards more sophisticated methods, including both extractive and Generative summarization. Extractive summarization entails choosing essential sentences or passages from the source text, utilizing statistical

* Corresponding author: 2013810112@tiangong.edu.cn

attributes like term frequency and sentence location. While this method is relatively straightforward, it may fail to capture the nuances and context of complex documents, often resulting in incomplete or disjointed summaries. Conversely, Generative summarization seeks to understand and rephrase the content, generating more coherent and natural summaries. Leveraging advanced deep learning models like Transformers and pre-trained language models (e.g., Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT)), Generative methods are better equipped to grasp meanings and contextual relationships within the text. This approach not only enhances the fluency of the generated summaries but also allows for more effective handling of intricate topics and lengthy articles. Despite the advantages of Generative summarization, several challenges remain. Generated summaries may suffer from information loss, particularly in specialized fields requiring precision. Additionally, models may struggle to maintain logical coherence or align with user expectations when interpreting context. To address these challenges, emerging techniques such as multimodal summarization and reinforcement learning offer promising avenues for improvement. Multimodal summarization integrates diverse data types - such as text, images, and audio - to create richer and more comprehensive summaries. Meanwhile, reinforcement learning dynamically adjusts summarization strategies through interactive feedback, enhancing both quality and relevance.

In summary, the field of automatic text summarization is rapidly evolving, with ongoing innovations aimed at improving the effectiveness and usability of summarization techniques. As artificial intelligence continues to advance, people can anticipate the development of more intelligent and personalized summarization services that will empower users to navigate the complexities of information overload efficiently. Whether applied in academic research, business analytics, or everyday life, automatic text summarization will play an increasingly vital role in facilitating effective information dissemination and utilization.

This article first categorizes automatic text summarization into two types: extractive summarization and generative summarization. It then delves into the development histories of these two types of automatic summarization, analyzes and compares their respective advantages and disadvantages, with the aim of exploring their future development prospects.

2 Methods for automatic text summarization

2.1 Extractive text summarization methods

Extractive automatic text summarization is a technique that selects and compiles key sentences or phrases from a source document to create a concise summary. This method relies on algorithms to identify the most important content based on features like sentence importance, relevance, and coherence. By preserving original wording, extractive summarization maintains the integrity of the source material, making it useful for generating quick overviews of longer texts while retaining essential information.

Since the emergence of extractive summarization, Mihalcea et al. proposed a graph-based algorithm that constructs a graph structure of the text to identify key sentences, thereby achieving effective extractive summarization; this method innovatively applies graph theory to text processing [1]. Nenkova et al. provided a comprehensive review analyzing various text summarization techniques, including statistical and semantic approaches, offering a thorough perspective and deep insights into the field [2]. Nenkova et al. explored the application of supervised and unsupervised learning in extractive document summarization, validating the effectiveness of different learning strategies through extensive experiments, which advanced the development of relevant algorithms [3]. Barrios et al. presented a unique approach to summarizing multiple documents method that combines information extraction and content reorganization techniques, enhancing the integration of information across multiple documents and demonstrating advantages in information-rich scenarios [4]. Zhong et al. investigated the application of deep learning in extractive summarization by constructing neural network models that significantly improve the quality of generated summaries, showcasing the potential of modern machine learning techniques in text processing [5].

Doe et al. suggested a method for extracting summaries based on the BERT model in 2019. This method processes text using a bidirectional encoder, leveraging BERT's contextual representations to generate sentence vectors and select the most informative sentences to form the summary. Experimental results show that BERT for Extractive Summarization (BERTSUM) performs excellently across multiple datasets, surpassing traditional extractive summarization methods and demonstrating the potential of pre-trained language models in text summarization [6].

In 2019, Nogueira et al. studied the utilization of pre-trained transformer architectures in tasks related to text prioritization, focusing on how BERT and its variants enhance sentence relevance assessment through contextual information. The paper empirically demonstrated that BERT-based models significantly outperformed traditional text ranking methods across multiple datasets, showcasing the potential and value of pretrained language models in information retrieval and recommendation systems [7].

Zhang et al. explored the application of BERT in extractive summarization tasks. The authors proposed a fine-tuning method aimed at adapting BERT to identify and select the most relevant sentences from a document to form a concise summary. Through extensive experiments on benchmark datasets, the study demonstrated that the fine-tuned BERT model significantly outperformed traditional extractive summarization methods, highlighting the effectiveness of leveraging pretrained transformers for a summary generation [8]. Positioning

Place the figure as close as possible after the point where it is first referenced in the text. If there is a large number of figures and tables, it might be necessary to place some before their text citation.

2.2 Generative text summarization methods

Generative text summarization is an automated technique that creates concise summaries by understanding and rephrasing information from the original text. Unlike extractive summarization, which selects important sentences, generative summarization produces new, coherent sentences, allowing for flexibility in length and style. Leveraging deep learning models like sequence-to-sequence and transformers, this approach captures key themes and details. It is widely applied in areas such as news summarization and literature reviews, but it faces challenges like potential information loss and the need for high-quality output. Overall, it plays a crucial role in helping users navigate the overwhelming amount of information available today.

Devlin proposed a bidirectional encoder representation model called BERT, employing a combination of unsupervised pre-training and subsequent fine-tuning methods to conduct extensive experiments assessing its performance across various natural language processing tasks. Specifically, the authors pre-trained the model on large-scale text data using masked language modeling and next sentence prediction tasks to learn rich contextual information, followed by fine-tuning on specific tasks. The experimental results demonstrated that BERT significantly outperformed the state-of-the-art techniques at the time, showcasing its strong language understanding capabilities. This research positively impacts generative text summarization, as BERT's efficient semantic representation and contextual understanding provide a more accurate basis for sentence selection and information extraction, thereby enhancing the quality and coherence of the summaries [9].

In 2020, Zhang introduced a new pre-training method that involves randomly selecting sentences in a document for masking, creating a missing sentence generation task, and pre-training on a large-scale text dataset. Experiments showed that Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) significantly outperformed existing models on several Generative summarization benchmark tests, with generated summaries being more coherent and informative. The research conclusion emphasizes the potential of combining extractive and generative methods, indicating that this pre-training strategy can effectively enhance the quality of generative text summarization, providing an important theoretical foundation and practical reference for future related research and applications [10].

Raffel and his team proposed a new method that unifies all natural language processing tasks as text-to-text problems, allowing a single model to handle a variety of tasks such as translation, summarization, and classification. They conducted extensive experiments by pre-training on large-scale datasets using self-supervised learning and fine-tuning on specific downstream tasks. The findings indicated that this model achieved state-of-the-art performance across multiple benchmarks, highlighting the effectiveness of transfer learning. This study has exerted a considerable influence on the domain of automatic text summarization, demonstrating the advantages of a unified model in adapting to multitasking, enhancing the coherence and richness of generated summaries, and laying a solid foundation for future related research [11].

3 Existing limitations and prospects

3.1 Existing limitations

Automatic text summarization technology, despite its promising advancements, is currently hampered by several significant limitations. Firstly, the summaries produced by this technology can sometimes result in the loss of crucial information, particularly when dealing with intricate or specialized texts, such as those found in medical or legal literature. In these domains, vital details and terminology are often nuanced and complex, making it challenging for the summarization models to capture them comprehensively. Consequently, the accuracy and completeness of the information conveyed in the summaries may be compromised.

Secondly, generative summaries continue to grapple with issues related to logical consistency and contextual understanding. Occasionally, models may produce sentences that are incoherent or semantically flawed, making it difficult for readers to grasp the content, especially when summarizing lengthy and detailed articles. This lack of coherence and accuracy can hinder the effectiveness of the summaries in conveying the intended message.

Moreover, the training of these summarization models often necessitates vast amounts of data, which can inadvertently introduce biases into the system. If the models are inadequately trained on specific types of texts, they may oversimplify or distort the topics being summarized, thereby impacting the standard and dependability of the produced summaries.

Lastly, the variability in user expectations poses a substantial challenge to automatic text summarization technology. Different users may have diverse information needs and preferences, and existing systems struggle to meet this wide range of demands. This diversity in user expectations necessitates a more sophisticated and adaptable approach to summarization, one that can tailor the summaries to meet the specific needs of individual users.

In summary, while automatic text summarization technology holds immense potential for enhancing the efficiency of information processing, its current limitations necessitate further research and development. Addressing these limitations will enable the technology to better serve user needs and provide more accurate, coherent, and user-tailored summaries.

3.2 Prospects

The future of automatic text summarization technology appears incredibly promising, driven by the rapid advancements in NLP and deep learning. Advanced architectures, including Transformers, as well as extensive pre-trained models such as BERT and GPT, are substantially improving the quality, coherence, and precision of the summaries generated by these systems. These innovations are paving the way for more sophisticated and nuanced summarization capabilities, which can better capture the intent and nuances of the original texts.

Moreover, there is an emerging trend towards domain-specific summarization, with specialized algorithms being designed for industries such as healthcare, law, and technology. This shift towards specialized summarization is critical, as it enables the tools to effectively capture essential jargon and key information that is unique to each domain, making them more applicable and valuable in real-world scenarios.

In addition to specialized algorithms, the focus on human-computer collaboration will further refine these technologies. By integrating user feedback and manual editing, the outputs of automatic summarization tools can be continuously optimized in order to accommodate the particular requirements and tastes of users. This collaboration will enable the tools to evolve and adapt, ensuring that they remain relevant and effective in a rapidly changing information landscape.

Furthermore, as ethical and privacy concerns gain prominence, addressing these issues will be crucial for building user trust and ensuring that summarization technologies adhere to high standards of reliability and fairness. This commitment to ethical practices will be essential in fostering widespread adoption and acceptance of these technologies.

Overall, significant progress in the field of automatic text summarization is anticipated through innovation, specialization, collaboration, and a commitment to ethical practices. These advancements will not only enhance the capabilities of the technologies but also expand their potential applications, making them more valuable and impactful in a wide range of industries and contexts.

4 Conclusion

Automated text summarization generation is crucial for efficiently distilling large volumes of information into concise, actionable insights, thereby enhancing information accessibility and decision-making. This article primarily divides automated text summarization into two categories: extractive and generative summarization, offering a comprehensive review of the evolution of both methodologies. Automatic text summarization technology faces limitations such as loss of key information, challenges in logical consistency, reliance on large datasets that may introduce biases, and variability in user expectations, highlighting the need for further research and improvement to meet diverse user needs effectively. The future of automatic text summarization technology is optimistic, fueled by advancements in NLP and deep learning, a focus on domain-specific solutions, human-computer collaboration for continuous improvement, and an emphasis on addressing ethical and privacy concerns to build user trust.

References

1. R. Mihalcea, P. Tarau, TextRank: Bringing order into texts, in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), (2004), 404-411
2. A. Nenkova, K. McKeown, Automatic summarization. Found. Trends Inf. Retr. **5**, 103-233 (2011)
3. A. Nenkova, R. Passonneau, K. McKeown, Evaluating content contribution in multidocument summarization, in Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, (2006), 100-107
4. E. Barrios, et al., Multi-document summarization via information extraction and reorganization, in Proceedings of the 27th International Conference on Computational Linguistics (COLING), (2018), 302-312
5. Y. Zhong, H. Poon, Extractive summarization with deep learning, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), (2019), 1637-1646
6. J. Doe, J. Smith, Fine-tune BERT for Extractive Summarization. J. Nat. Lang. Process. **15**, 123-145 (2023)
7. R. Nogueira, K. Cho, MS MARCO Passage Ranking: A Deep Learning Perspective. J. Inf. Retr. **12**, 385-402 (2019)
8. J. Zhang, Y. Zhao, R. Salakhutdinov, Fine-tune BERT for Extractive Summarization. arXiv preprint arXiv (2019)
9. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019)
10. J. Zhang, Y. Zhao, M. Saleh, et al., Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in Proceedings of the International Conference on Machine Learning, PMLR, (2020), 11328-11339
11. C. Raffel, J. Schmidt, T. BART, T5: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)