RESEACH ON TEXT SUMMARIZATION AND SOME EFFECTIVE SUMMARIZATION METHODS

Trần Thị Thu Phương, Nguyễn Quốc Tuấn
Hanoi Metropolitan University
Lê Thị Hằng
Center for Equipment Management and Science and Technology
Nguyễn Xuân Nhi
IT D2021B

Abstract: *This study examines extractive and abstractive text summarization, emphasizing deep learning techniques such as Transformer architectures and reinforcement learning. Extractive summarization selects key sentences, while abstractive summarization generates human-like summaries by rephrasing content. Key datasets like Vietnews and Wikilingua are highlighted for their role in training models for low-resource languages like Vietnamese. The research addresses challenges in maintaining coherence and semantic accuracy, proposing solutions to enhance summarization quality. Future directions include improving evaluation metrics, refining coherence in Vietnamese summaries, and advancing multilingual models. By integrating modern techniques and addressing key challenges, this study contributes to the development of more accurate and reliable automatic summarization systems.*
Keywords*: Datasets, deep learning, model, reinforcement learning, text summarization****.***

Nhận bài ngày 13.02.2025; gửi phản biện, chỉnh sửa, duyệt đăng ngày 20.03.2025
Liên hệ tác giả: Trần Thị Thu Phương; email: tttphuong2@daihocthudo.edu.vn

1. INTRODUCTION

Text summarization is the process of condensing information from a long document into a shorter summary that retains the core meaning and key information of the original text. With the rapid growth of digital data and the need for quick access to information, automatic summarization systems have become valuable tools for users. Among these, two main summarization methods are widely used: extractive summarization and abstractive summarization. The example below illustrates the text and its summarization.

Source text: *The research report on one-time social insurance (SI) withdrawal in Vietnam: Trends, Challenges, and Recommendations, recently published by the ILO and WB, reveals that one-time insurance payouts account for a significant proportion of all one-time withdrawals in Vietnam, rising from 82% during the 2013–2016 period to 93% in the 2016–2019 period. In 2019, approximately 69% of these one-time payouts were made to female workers under the age of 35. These women often require the funds to cover expenses for childbirth and child-rearing.*

The ILO assesses that while one-time SI withdrawals may appear substantial and appealing to workers, they pose several challenges. No one can predict how long they will live after retirement—whether 5 years or 30 years—nor how much they will need to spend over the course of their lifetime. Without proper savings plans, workers face significant financial difficulties in old age.

Many individuals use the withdrawn funds for business investments, purchasing new homes, funding their children's overseas education, or traveling abroad. However, most quickly deplete the money, even those who have carefully devised financial plans.

The ILO cites research conducted in Malaysia during the 2000s, which showed that most workers who withdrew one-time insurance payouts for early retirement spent the entire sum within three years. Ultimately, they had to rely on government-provided social assistance programs for the poor. This scenario imposes a financial burden on society, including those who are actively paying taxes.

The summarized text: *The International Labour Organization (ILO) in Vietnam highlights that most workers quickly spend the lump sum withdrawn from social insurance (SI) and face difficulties in old age.*

Extractive summarization selects and combines key sentences or paragraphs from the original document to create a summary. In contrast, abstractive summarization requires the model to rephrase the main ideas of the document using new wording, closely resembling human summarization. Both methods have their strengths and limitations, depending on the characteristics of the source material and the intended use of the automatic summarization system.

This paper aims to analyze and compare various text summarization methods, particularly advanced deep learning techniques such as Transformer and reinforcement learning. We also evaluate popular Vietnamese and international summarization datasets, highlight challenges, and discuss solutions to improve the quality of text summarization, especially in terms of semantic consistency and accuracy.

The remainder of this paper is structured as follows. Section 1: Introduction provides an overview of text summarization, emphasizing extractive and abstractive methods while outlining the research objectives. Section 2: Content explores text summarization techniques, key datasets, advanced approaches such as Transformer-based models and reinforcement learning, and proposed solutions to enhance summary quality. Finally, Section 3: Conclusion summarizes key findings and suggests future research directions to further improve automatic text summarization

2. CONTENT

2.1. Text Summarization Methods

Text summarization techniques have evolved from rule-based and statistical methods to modern deep learning models. This section provides a detailed analysis of two main summarization methods and recent advancements in each approach.

2.1.2. Extractive summarization

The extractive summarization method selects sentences or paragraphs with high importance from the original text to construct a summary. Traditional extraction techniques often use statistical indicators such as word frequency, sentence position, or the weight of important keywords to identify the most meaningful sentences.

Example with original text:

"*Education is a critical factor in human and societal development. It provides us with the knowledge, skills, and values needed to become responsible citizens capable of contributing positively to societal progress. The education system must continuously improve and adapt to meet the changing demands of the modern world. Investing in education is not only an investment in individuals but also an investment in the future of society as a whole.*"

Extractive summary:

"Education is a critical factor in human and societal development. The education system must continuously improve and adapt to meet the changing demands of the modern

world. Investing in education is not only an investment in individuals but also an investment in the future of society as a whole."

2.1.3.Abstractive summarization

Unlike extractive summarization, abstractive summarization requires the model to understand and rephrase the content of the document in a natural, concise, and new way. This approach employs complex language models such as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and especially Transformer models, which enable the system to analyze context and interpret complex information.

Example:

For the same original text, an abstractive summary might be:

"*Education plays a vital role in human and societal development, providing essential knowledge and skills. The education system must adapt to global changes, and investment in education is an investment in the community's future*."

TABLE 1 FOCUSES ON COMPARING THE STRENGTHS AND WEAKNESSES OF EXTRACTIVE AND ABSTRACTIVE SUMMARIZATION. ADDITIONALLY, IT PROVIDES USEFUL USE CASES FOR EACH METHOD.

TABLE 1 COMPARISON OF EXTRACTIVE VS. ABSTRACTIVE SUMMARIZATION

Summarization Method	Strengths	Weaknesses	Best Use Cases
Extractive Summarization	Preserves original sentences, maintains factual accuracy, computationally efficient	Lacks fluency, may not always form a coherent summary, depends on extracted sentences	When maintaining original wording is crucial, for legal and technical documents
Abstractive Summarization	Generates natural, human-like summaries, capable of paraphrasing and generalizing information	Requires complex models, higher computational cost, risk of generating inaccurate or hallucinated	For producing more readable and concise summaries, useful in news, academic, and conversational AI applications

2.2. Text summarization datasets

Datasets are a crucial component in training and evaluating the performance of summarization models. High-quality datasets help models understand the structure, semantics, and context of a document.

2.2.1 International datasets

Popular international datasets for text summarization research provide essential resources for developing and evaluating model performance across various types of content. The DUC-2004[1] dataset, comprising 500 articles from reputable sources like the New York Times and Associated Press, includes four human-written reference summaries per article, establishing it as a benchmark for evaluating automatic summarization models. Gigaword [1], with around 4 million headline-article pairs, offers a diverse range of sources and language styles, making it highly valuable for building effective news summarization models. Another widely used dataset, CNN/DailyMail [2], contains over 300,000 articles and is specifically useful for training abstractive summarization models on content with complex sentence structures. Lastly, XSum [3], a BBC-sourced dataset with 226,000 articles, each paired with a concise one-sentence summary, is particularly suited for models

[1] https://duc.nist.gov/duc2004/

that aim to capture the core idea of a document. Together, these datasets contribute significantly to advancements in summarization model training and evaluation.

2.2.2. Vietnamese datasets

In the field of Natural Language Processing (NLP), high-quality datasets are essential for developing effective models, especially for low-resource languages like Vietnamese. While Vietnamese NLP resources are still limited, there are notable datasets that support research and implementation of automatic summarization models.

Wikilingua [4]: This multilingual dataset includes pairs of source and summary texts in multiple languages, including Vietnamese. Wikilingua's documents are often reference materials with considerable length and complexity, suitable for research on multilingual summarization models and applications for detailed texts. It also supports cross-lingual learning by leveraging knowledge from one language to improve model performance in another.

Vietnews [5]: A large-scale Vietnamese dataset with 150,704 document-summary pairs, primarily sourced from news articles covering various topics such as politics, society, sports, and culture. This dataset is valuable in helping models learn to identify and summarize information in different contexts, creating concise summaries while retaining core content. Vietnews is especially important in Vietnamese research, where high-quality NLP resources are still limited.

Zalo AI challenge summarization dataset [6]: Developed within the framework of the Zalo AI Challenge, this dataset includes human-written document-summary pairs focused on current events and technology topics. This dataset is highly applicable in researching and developing automatic summarization models and provides a basis for researchers to experiment and compare the effectiveness of summarization algorithms for Vietnamese texts.

The Vietnamese datasets above enable models to understand Vietnamese language characteristics and semantic structures, improving the accuracy and consistency of summaries. However, challenges remain, particularly in maintaining semantic coherence and accuracy in complex documents, requiring models to handle long and intricate contexts.

2.3. Advanced text summarization techniques

Modern text summarization techniques have made significant advances due to deep learning models, particularly Transformer models and reinforcement learning. These techniques enhance the model's ability to process semantics and improve the coherence and naturalness of summaries.

2.3.1. Transformer-based models

BERTSum [7]: A variant of BERT developed for extractive summarization, BERTSum employs a two-step approach involving encoding and decoding, where encoding helps identify critical sentences within the original document. This method enables summarization systems to better understand the structure and semantics of the source text, resulting in more coherent summaries. BERTSum has proven effective for long documents such as articles and detailed reports.

BART [8]: BART (Bidirectional and Auto-Regressive Transformers) uses both an encoder and a decoder based on Transformer, effectively performing abstractive summarization. BART excels in text transformation tasks, including machine translation, summarization, and text generation. With its context encoding capability, BART generates summaries that are both natural and readable.

PEGASUS [9]: PEGASUS is a specialized encoder-decoder model for abstractive summarization, trained to predict masked sentences within the source document,

simulating the summarization process. By masking critical sentences and reconstructing them, PEGASUS achieves high performance on concise summarization datasets such as XSum and CNN/DailyMail, where brief yet semantically coherent summaries are required.

2.3.2. ViT5 - Vietnamese summarization model

ViT5 [10] is a language model tailored for Vietnamese, based on the T5 (Text-To-Text Transfer Transformer) architecture. With an encoder-decoder structure, ViT5 effectively handles automatic summarization tasks, especially in Vietnamese contexts. The model, trained on Vietnamese monolingual datasets, demonstrates strong contextual understanding and grammatical structure recognition, enhancing the performance of NLP systems for Vietnamese..

2.3.3. Reinforcement learning (RL)

In text summarization, reinforcement learning is used to optimize summaries based on human feedback. RL allows models to adjust their summarization strategies based on user evaluations of accuracy and readability. One popular RL method is PPO (Proximal Policy Optimization) [11], enabling models to generate high-accuracy summaries tailored to user requirements. Studies show that RL-applied models yield significant improvements in the quality and coherence of summaries.

2.4. Solutions for improving summary quality

Improving the quality of automatic text summaries is a major challenge in NLP research, particularly as models must ensure that summaries are both accurate and natural. This section presents specific solutions to enhance consistency, semantic accuracy, and information conveyance in summaries.

2.4.1. Evaluating summary quality

Evaluation is a crucial step to ensure that automatic summaries meet semantic and accuracy requirements. Evaluation can be conducted using automatic metrics or manual assessment by human evaluators. Manual evaluation is reliable due to human comprehension of semantics and context. Automatic evaluation using metrics such as ROUGE, BERTScore enables quick, consistent measurements suitable for large-scale tasks.

ROUGE Score: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [12] is the most commonly used metric for automatic summarization evaluation. ROUGE-N measures the n-gram overlap between the generated summary and the reference, while ROUGE-L measures the longest common subsequence (LCS), ensuring sentence coherence. ROUGE is valuable for short, direct summarization tasks, though it primarily reflects word overlap and may lack full semantic coherence.

BERTScore [13]: Based on BERT, BERTScore compares semantic representations of the generated and reference summaries. BERTScore can assess semantic consistency and context accuracy, supplementing ROUGE's limitations. This metric is particularly suitable for summaries in diverse, complex languages like Vietnamese, where coherence and semantics are crucial.

2.4.2. Model optimization through Fine-tuning

Fine-tuning is a process in which a model's parameters are adjusted on specific datasets to significantly improve the accuracy and quality of generated summaries. This process is particularly valuable when applied to specialized datasets tailored to capture unique domain-specific and linguistic characteristics. For example, when fine-tuning on domain-specific summarization datasets, the model learns to identify key informational elements unique to specific fields, enabling it to produce summaries that maintain high semantic accuracy and relevance across varying contexts. This approach enhances the model's understanding of specialized vocabulary, concepts, and structural nuances within those

fields, ultimately improving the semantic precision of its output. Additionally, fine-tuning on multilingual and complex-context datasets is essential for enabling the model to handle multiple languages and adapt to the syntactic and grammatical nuances characteristic of each language, including Vietnamese. By learning to accommodate these linguistic variations, the model enhances its coherence and is better equipped to convey detailed, contextually appropriate information within summaries. This approach is particularly effective for complex documents with intricate informational structures, allowing the model to provide accurate, concise, and coherent summaries across a range of linguistic and contextual demands.

2.4.3. Enhancing consistency through reinforcement learning

Reinforcement Learning has been applied to improve the quality of text summarization by incorporating human feedback [14]. The process of applying reinforcement learning in text summarization involves the following steps:

Step 1. Collecting human feedback: The summary generated by the model is evaluated by users based on criteria such as semantic accuracy, coherence, and information conveyance. User feedback provides the model with information on the necessary improvements needed to achieve better quality.

Step 2. Training the reward model: Based on user feedback, the reward model learns to identify the characteristics of a high-quality summary. Specifically, this model uses evaluation criteria such as coherence and semantic accuracy to assign scores to the summaries. Summaries that exhibit higher semantic accuracy and coherence are assigned higher reward scores.

Step 3. Optimizing the summarization policy with PPO: PPO is an advanced reinforcement learning algorithm used to adjust the summarization model's behavior based on feedback from the reward model. By optimizing the summarization policy, the model can generate higher-quality summaries that meet user expectations in terms of accuracy and readability. Research results show that summarization models using RL with PPO have the ability to maintain context, semantics, and enhance the consistency of the summary.

2.4.4 Semantic enhancement with transfer learning

Transfer learning [15] is a powerful approach that allows models to leverage knowledge acquired from a large and diverse dataset and apply it effectively to tasks with limited data resources. By pre-training on a large dataset, the model gains a deep understanding of language semantics, structure, and contextual relationships, which it can then transfer to specialized tasks with smaller datasets. For Vietnamese text summarization, transfer learning is particularly valuable, as it enables models to first learn from extensive, semantically rich datasets in larger, high-resource languages, or multilingual datasets, before fine-tuning on more focused Vietnamese datasets.

This approach significantly enhances the model's ability to capture nuanced semantics and context in Vietnamese texts. The pre-trained model, having already developed a strong linguistic foundation, can more effectively understand and represent complex ideas, idiomatic expressions, and context-dependent meanings in Vietnamese. This is especially beneficial when dealing with specialized or intricate texts that require a nuanced understanding of language, as transfer learning helps the model adapt and generalize its knowledge to accurately capture the core meaning, relationships, and subtleties within Vietnamese-language summaries. Ultimately, transfer learning improves the quality of summarization by ensuring that the output retains both coherence and contextual relevance, even when working with smaller and potentially less comprehensive Vietnamese datasets.

2.4.5. Combining extractive and abstractive summarization

Finally, a proposed solution is to combine extractive and abstractive summarization to produce highly accurate summaries. This approach first uses extractive summarization to select important sentences from the text, then applies abstractive summarization to restructure and rephrase the content. This combination helps the summary maintain coherence while retaining the semantic accuracy of the original document, making it suitable for long and complex texts.

A study by Liu et al demonstrated the effectiveness of this approach by applying a hybrid extractive-abstractive model to summarize lengthy scientific documents [16]. The study first employed extractive summarization to identify key sentences, ensuring factual accuracy, and then used a Transformer-based abstractive model to generate a more concise and coherent summary. The results showed that this combined method significantly improved summary quality, as evaluated using ROUGE scores, compared to using either approach alone. Similarly, another study by Gehrmann et al proposed a unified model that integrates extractive and abstractive techniques through a hierarchical attention mechanism, achieving state-of-the-art results on the CNN/Daily Mail dataset [17]. These findings highlight the advantages of combining both methods, making it a promising solution for summarizing complex and information-dense texts.

3. CONCLUSION

This study has analyzed automatic text summarization methods, from traditional methods like extractive summarization to advanced deep learning techniques such as Transformer and reinforcement learning. We also introduced popular Vietnamese and international datasets, particularly the Vietnews and Wikilingua datasets, which contribute to improving Vietnamese summarization research. While modern techniques have improved coherence and semantic accuracy in text summarization, challenges remain regarding consistency and accuracy, especially for complex documents.

Future research in text summarization may focus on several key areas to further improve model performance and applicability. Developing new semantic evaluation metrics is essential, as current metrics like ROUGE primarily assess word overlap without fully capturing semantic depth; new metrics could provide more accurate assessments of summarization quality. Additionally, enhancing coherence for Vietnamese summaries is crucial, as existing models often struggle with maintaining semantic connections in lengthy texts; research could address this by improving semantic linkage across paragraphs, resulting in more coherent Vietnamese summaries. Another promising direction is the development of multilingual models, which would broaden the utility of summarization systems across languages, especially benefiting low-resource languages through cross-lingual techniques. Lastly, applying reinforcement learning to complex document types, such as dialogues and specialized content, could significantly expand the practical applications of summarization models across diverse fields like education, healthcare, and media, allowing for more tailored and context-sensitive summaries.

REFERENCES

1. Courtney Napoles et al (2012), *Annotated Gigaword*, In Proceed-ings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowlege Extraction, pp. 95-100.
2. Nallapati et al (2016), *Abstractive text summarization using sequence-to-sequence RNNs and beyond,* In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pp. 280-290.
3. Shashi Narayan et al (2018), *Don't give me the details, just the summary! topic-aware*

convolutional neural networks for extreme summarization, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, p. 1797–1807, October-November .

4. "A Multilingual Abstractive Summarization Dataset," [Online]. Available: https://github.com/esdurmus/Wikilingua. [Accessed 12 9 2024].
5. Nguyen et al(2019), *VNDS: A Vietnamese Dataset for Summarization*,In: 6th NAFOSTED Conference on Information and Computer Science (NICS).
6. "Zalo AI Challenge 2020: News Summarization - Runner-up solution," [Online]. Available: https://github.com/btcnhung1299/zaloai-2020-news-summarization. [Accessed 11 11 2024].
7. Lapata et al (2019), *Text summarization with pretrained encoders*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing , p. 3730–3740.
8. Mike Lewis et al (2020),*BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension*, In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 7871–7880.
9. Jingqing Zhang et al (2020), *PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization*, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, p. 11328–11339.
10. "ViT5," . [Online]. Available: https://github.com/vietai/ViT5 [Accessed 11 11 2024]
11. "Proximal Policy Optimization (PPO)," [Online]. Available: https://huggingface.co/learn/deep-rl-course/unit8/introduction. [Accessed 11 2024].
12. C.-Y. Lin (2004), *Rouge: A package for automatic evaluation of summaries*., Text summarization branches out.
13. Tianyi Zhang et al (2020), *BERTScore: Evaluating Text Generation with BERT*, ICLR2020.
14. Stiennon et al (2020), *Learning to summarize from human feedback*, In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Proces., p. 3008–3021.
15. T. S. e. al (2024), *Transfer Learning for Finetuning Large Language Models*, ArXiv.org.
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2020), *Text summarization with pretrained encoders*, *arXiv preprint*, arXiv:2006.05354.
17. Gehrmann, S., Deng, Y., Rush, A. M. (2018), *Bottom-up abstractive summarization*, *arXiv preprint*, arXiv:1808.10792.

NGHIÊN CỨU VỀ TÓM TẮT VĂN BẢN VÀ MỘT SỐ PHƯƠNG PHÁP TÓM TẮT HIỆU QUẢ

Tóm tắt: *Nghiên cứu này xem xét các phương pháp tóm tắt văn bản trích xuất và tóm tắt khái quát, tập trung vào các kỹ thuật học sâu như kiến trúc Transformer và học tăng cường. Tóm tắt trích xuất lựa chọn các câu quan trọng, trong khi tóm tắt khái quát tạo ra bản tóm tắt giống con người bằng cách diễn đạt lại nội dung. Các bộ dữ liệu quan trọng như Vietnews và Wikilingua được nhấn mạnh vì vai trò của chúng trong việc huấn luyện mô hình cho các ngôn ngữ có ít tài nguyên như tiếng Việt. Nghiên cứu cũng đề cập đến các thách thức trong việc duy trì tính mạch lạc và chính xác ngữ nghĩa, đồng thời đề xuất các giải pháp để nâng cao chất lượng tóm tắt. Định hướng tương lai bao gồm cải thiện các tiêu chí đánh giá, tối ưu hóa tính mạch lạc trong tóm tắt tiếng Việt và phát triển các mô hình đa ngôn ngữ tiên tiến. Bằng cách tích hợp các kỹ thuật hiện đại và giải quyết các thách thức quan trọng, nghiên cứu này đóng góp vào sự phát triển của các hệ thống tóm tắt tự động chính xác và đáng tin cậy hơn.*

Từ khóa*: Bộ dữ liệu, học sâu, học tăng cường, mô hình, tóm tắt văn bản.*