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/