Literature Review: Automatic Text Summarization of News Articles Using Natural Language Processing

Noor Magdy 16005059 - Mohamed Wael 19011272
M.Sc. in Data Science, Course: INCS 903 Natural Language Processing

Abstract— **Automatic text summarization plays an important role in natural language processing by enabling concise representation of long news documents. This project focuses on the summarization of television and environmental news content. The report reviews existing extractive, abstractive, and neural summarization approaches, with particular emphasis on transformer-based models for news data. In addition, an exploratory analysis of a news transcript dataset is conducted to examine its textual characteristics and challenges relevant to summarization. The insights gained provide a foundation for subsequent modeling and evaluation stages of the project.**

I. Introduction

The fast expansion of digital news and online textual information has made automatic text summarization a major area of study in natural language processing. Generating succinct representations of lengthy documents while retaining their most crucial information is the aim of summarization systems. Effective summary is especially useful when it comes to news items because readers frequently want to quickly access important information and ideas without having to read the full article.

With an emphasis on news-related content, this section examines previous studies on artificial text summarizing. Traditional and neural models, extractive and conceptual summarizing strategies, and more modern transformer-based methods that have proven to perform well on news summary tasks are all covered.

II. Overview of Automatic Text Summarization

Automatic text summarization is a core task in natural language processing that aims to condense long documents into shorter, informative summaries. With the increasing volume of digital news content, summarization systems have become essential for enabling efficient information access. Modern summarization research focuses primarily on data-driven approaches, particularly deep learning models that can capture semantic meaning and contextual relationships in text.

This literature review focuses on automatic summarization methods applied to news content. The following sections discuss extractive and abstractive approaches, neural sequence-to-sequence models, and recent transformer-based techniques that have demonstrated strong performance in news summarization tasks.

III. Extractive and Abstractive Summarization Approaches

Text summarization methods are commonly categorized into **extractive** and **abstractive** approaches. Extractive summarization constructs a summary by selecting and ranking the most important sentences from the original document, while abstractive summarization generates new sentences that may paraphrase or compress the source content. Extractive approaches are generally easier to implement and interpret, but they often suffer from redundancy and limited coherence, especially in longer news articles.

Several survey studies highlight the fundamental differences between these two paradigms and analyze their respective advantages and limitations. Reviews of neural summarization research indicate that extractive methods rely heavily on sentence importance estimation, whereas abstractive approaches require stronger language generation capabilities and semantic understanding of the document content [6], [7]. While extractive methods are more robust and less prone to grammatical errors, abstractive models are better suited for producing fluent and human-like summaries.

Recent studies focusing on deep learning-based summarization emphasize that the shift toward abstractive methods has been driven by advances in neural architectures and training techniques [1], [9] However, these studies also note that abstractive summarization introduces challenges such as content hallucination, factual inconsistency, and difficulty in identifying the most salient information.

IV. Neural Network-Based Summarization Using Sequence-to-Sequence Models

The introduction of neural networks marked a major advancement in automatic text summarization research. Sequence-to-sequence (Seq2Seq) models, originally proposed for machine translation tasks, were later adapted for abstractive text summarization. These models typically consist of an encoder that transforms the input document into a latent representation and a decoder that generates the summary word by word.

Early work demonstrated that Seq2Seq models based on recurrent neural networks (RNNs) are capable of learning meaningful representations of textual data and generating coherent summaries [8]. The incorporation of attention mechanisms further improved performance by allowing the model to focus on relevant parts of the input during summary generation. This development enabled abstractive summarization systems to produce more informative outputs compared to traditional extractive approaches.

Despite these improvements, subsequent studies revealed several limitations of RNN-based Seq2Seq models [12]. In particular, these models struggle with long input documents, often producing repetitive content or omitting important information. Such limitations motivated further research into alternative architectures that can better capture long-range dependencies in text.

V. Transformer-Based Models for News Summarization

More recent research in text summarization has shifted toward **transformer-based architectures**, which rely on self-attention mechanisms instead of recurrent structures. Transformers have demonstrated strong performance across various natural language processing tasks due to their ability to model long-range dependencies and process text in parallel.

Several studies show that transformer-based models achieve state-of-the-art results in both extractive and abstractive news summarization tasks [2], [5]. These models are particularly effective in handling the complexity and length of news articles, making them well-suited for large-scale summarization applications. Research on automated news summarization further highlights the advantages of transformers in generating concise and coherent summaries while maintaining contextual relevance [10], [11].

In addition to purely extractive or abstractive approaches, recent work has explored hybrid models that combine both paradigms within a unified framework. Such approaches aim to improve content selection while preserving fluent text generation [3]. These findings suggest that transformer-based models currently represent the most effective solution for news summarization, while still leaving room for improvement in areas such as factual consistency and domain adaptation.

VI. Summary of Literature Findings

Overall, the reviewed literature demonstrates a clear progression in automatic text summarization research, from extractive techniques to neural and transformer-based approaches. Survey studies consistently report that while extractive methods remain reliable and efficient, abstractive and hybrid models offer greater flexibility and more natural summaries when supported by advanced neural architectures [6], [9].

However, existing research also highlights persistent challenges, including effective content selection, redundancy reduction, and handling varying document lengths in news data [1], [7]. These challenges motivate continued investigation into summarization techniques tailored to news content, providing the foundation for the work presented in this project.

VII. Implications for This Project

The reviewed literature demonstrates that automatic text summarization has evolved significantly from traditional extractive techniques toward neural and transformer-based approaches, particularly in the context of news content. Survey studies consistently indicate that while extractive methods offer robustness and simplicity, abstractive and hybrid models are more effective at producing fluent and informative summaries when supported by advanced neural architectures [6], [9].

Research on sequence-to-sequence models highlights the potential of neural networks to generate abstractive summaries, but also reveals challenges related to long input texts, redundancy, and content selection [8], [12]. More recent transformer-based models address many of these limitations by leveraging self-attention mechanisms, enabling better handling of long-range dependencies and improved performance on news summarization tasks [2], [5]. However, existing studies also emphasize that summarization performance remains sensitive to data characteristics such as document length, domain specificity, and textual structure [1], [7].

Motivated by these findings, this project focuses on the summarization of news-related textual content, aiming to explore suitable summarization approaches for television and environmental news data. By building on insights from prior work, the project seeks to examine how modern summarization techniques can be applied to domain-specific news datasets with varying text lengths and contextual complexity.

References

[1] Sanjrani, A. A., Saqib, M., Rehman, S., & Ahmad, M. S. (2024). Text summarization using deep learning: a study on automatic summarization. *The Asian Bulletin of Big Data Management*, *4*(4), 216-226.

[2] Gupta, A., Chugh, D., Anjum, & Katarya, R. (2022). Automated news summarization using transformers. In *Sustainable advanced computing: select proceedings of ICSAC 2021* (pp. 249-259). Singapore: Springer Singapore.

[3] Hsu, W. T., Lin, C. K., Lee, M. Y., Min, K., Tang, J., & Sun, M. (2018). A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

[4] Kedzie, C., McKeown, K., & Daume III, H. (2018). Content selection in deep learning models of summarization. *arXiv preprint arXiv:1810.12343*.

[5] Pilault, J., Li, R., Subramanian, S., & Pal, C. (2020, November). On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 9308-9319).

[6] Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, *9*, 13248-13265.

[7] Giarelis, N., Mastrokostas, C., & Karacapilidis, N. (2023). Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*, *13*(13), 7620.

[8] Nallapati, R., Zhou, B., Dos Santos, C., Gulçehre, Ç., & Xiang, B. (2016, August). 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).

[9] Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. *Mathematical problems in engineering*, *2020*(1), 9365340.

[10] Singh, R. K., Khetarpaul, S., Gorantla, R., & Allada, S. G. (2021). SHEG: summarization and headline generation of news articles using deep learning. *Neural Computing and Applications*, *33*(8), 3251-3265.

[11] Rani Krishna, K. M., Somasundaram, K., Arulmozhivarman, P., Immanuel, S. A., & Rajkumar, E. R. (2025). Deep learning for text summarization using NLP for automated news digest. *Scientific Reports*, *15*(1), 36343.

[12] Nallapati, R., Xiang, B., & Zhou, B. (2016). Sequence-to-sequence rnns for text summarization.