STM JOURNALS

Current Trends in
Signal Processing

ISSN: 2277-6176 (Online)
ISSN: 2321-4252 (Print)
Volume 15, Issue 1, 2025
January–April
DOI (Journal): 10.37591/CTSP

https://journals.stmjournals.com/ctsp

Review CTSP

Exploring Technologies for Extractive Text Summarization: A Review of Transformer and Reinforcement Learning Models

Rajshree Tarapure*, Sharayu Mulay, Ashlesha Kathe,
Prapti Jadhav, Shalaka Deore

Abstract

In recent years, the size of information on the Internet has increased exponentially. Therefore, a solution is needed to transform large amounts of raw data into useful information that the human brain can understand. Automatic Text Summarization (ATS) is a part of Natural Language Processing (NLP) that aims to take long texts and shorten them, keeping the most important information in a clear and easy-to-understand way. This research report explores methods for extracting content from text and its relevance. Reinforcement Learning is used in extractive summarization to make summaries more relevant, clear, and varied. It helps the system learn to create better summaries by rewarding them when the summary meets these goals. According to the review, Transformer models excel in determining how words in a sentence relate to one another as well as the overall meaning. This skill has shown to be quite useful for tasks such as text summarization. They excel at navigating extensive chunks of literature and extracting the most important ideas. This way, the system gets better at picking the right sentences and organizing them effectively. We delve into the Transformer model, the implementation of BERT, and the integration of support learning. This study describes the main processes operating in content extraction and their important role in improving the quality and performance of text collection.

Keywords: Extractive, single document, transformer model, reinforcement learning, ATS, NLP

INTRODUCTION

In an era characterized by the relentless proliferation of digital information, the ability to distill vast volumes of text into concise, informative summaries has become an indispensable tool for both individuals and organizations. Extractive text summarization, a subfield of natural language processing, serves as the conduit through which the deluge of textual data is transformed into manageable and coherent insights.

***Author for Correspondence**
Rajshree Tarapure
E-mail: rajshreetarapure@gmail.com

Student, Department of Computer Engineering, Modern Education Society's College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

Received Date: December 07, 2024
Accepted Date: January 06, 2025
Published Date: January 16, 2025

Citation: Rajshree Tarapure, Sharayu Mulay, Ashlesha Kathe, Prapti Jadhav, Shalaka Deore. Exploring Technologies for Extractive Text Summarization: A Review of Transformer and Reinforcement Learning Models. Current Trends in Signal Processing. 2025; 15(1): 1–6p.

This project embarks on an exploration at the intersection of cutting-edge technology and the art of language comprehension. Our endeavor is to employ the formidable capabilities of transformer models, exemplified by the likes of BERT, in combination with the dynamic decision-making process of reinforcement learning (RL), to revolutionize the domain of text summarization.

The journey begins with the comprehensive encoding of sentences from single documents using transformer models, which are celebrated for their ability to grasp context, semantics, and meaning.

These encoded representations become the foundation upon which we build our vision of efficient and informative summarization [1–4].

RELATED WORK

Gao *et al*. explored L2R and RL-based summarization approaches, with a specific focus on RELIS. Their findings indicated that RELIS significantly outperformed RL-based models and achieved competitive performance compared to state-of-the-art neural summarizers [5]. Notably, RELIS required significantly less training time and data for training, making it an efficient choice for summarization tasks.

Ranganathan and Abuka conducted an evaluation of their model using the ROUGE metrics, which serve as valuable tools for assessing the quality of model-generated summaries by comparing them with human-generated summaries [6]. The evaluation focused on ROUGE-1, ROUGE-2, and ROUGE-L scores. To enhance their evaluation, the study included manual summaries of the drug review dataset. However, the results from the UCI drug reviews dataset were less impressive than those from the BBC news dataset, primarily due to the limited number of training samples. Specifically, the ROUGE-1, ROUGE-2, and ROUGE-L scores were reported as 69.05, 59.70, and 52.97, respectively. This highlighted the challenges of using transformer models like BERT, which demand significant computational resources and may not be accessible for resource-constrained applications.

Srikant *et al*. employed the BERT model to generate contextual embeddings for each sentence within the document using BERT-Based Feature Extraction [7]. The research paper presented experimental results that showcased the effectiveness of dynamic clustering and co-reference resolution techniques in combination with BERT embeddings. The ROUGE-1(F1), ROUGE-2(F1), and ROUGE-L(F1) scores reported in this context were 41.4, 17.9, and 37.9, respectively, demonstrating the potential of BERT embeddings in text summarization.

Adhikari discussed various summarization methods, including Query based, Structured based, and Semantic based approaches [8]. They highlighted the common usage of metrics like Rouge and TFIDF and suggested the potential of GANs and transfer learning for more precise summaries.

Jugran *et al*. focused on the Extractive Approach and utilized NLP tools such as SpaCy, CoreNLP, and NLTK. Their work compared these tools in terms of Precision, Recall, and F-Score [9]. Results indicated that Core NLP outperformed the others, achieving high scores in all three metrics.

Akhmetov *et al*. performed the optimization of ROUGE1 scores using the VNS heuristic algorithm [10]. Their research obtained ROUGE1 scores of 0.55 for both techniques and ROUGE2 scores of 0.21 and 0.23. In contrast, sophisticated neural network architectures achieved higher ROUGE1 and ROUGE2 scores on the arXive dataset.

In our quest to redefine the landscape of extractive single document text summarization, we embark on a journey that combines the dynamic capabilities of transformer models, inspired by Srikanth [7], with the decision-making process of reinforcement learning, following the footsteps of Gao [5]. While these base papers have laid the foundation for transformative techniques in their respective domains, our vision extends further. We aspire to leverage their innovations and adapt them uniquely to the realm of single document summarization, aiming not only to achieve higher scores but also to elevate the quality and relevance of the summaries generated. In doing so, we seek to push the boundaries of what is possible and set new standards in the field of text summarization.

METHODOLOGY

Dataset

In our current exploration of the CNN Daily Mail dataset, our research is exclusively focused on this rich source of news articles. The CNN Daily Mail dataset provides a comprehensive collection of news

articles paired with high-quality summaries, making it an ideal choice for our extractive text summarization endeavors. By concentrating solely on the CNN Daily Mail dataset, we aim to maximize our understanding of its nuances and intricacies, allowing us to develop more effective summarization models tailored specifically to this domain. This dedicated focus on a single dataset enables us to delve deeply into its characteristics, enhancing our model's ability to extract salient information and generate coherent summaries across various topics and document types.

Variables Used

- *Text content*: Represents the main body of news articles extracted from the CNN Daily Mail dataset.
- *Summary*: Ground truth summaries accompanying each news article, providing concise representations of key points.
- *Metadata*: Additional information such as publication date, article ID, and source may be utilized for contextual analysis or preprocessing steps.

Data Collection and Preprocessing

For this project, our primary dataset consists of articles from the CNN Daily Mail dataset, a comprehensive collection of news articles paired with summaries. These articles cover a wide range of topics and provide substantial textual content for our analysis. In the preprocessing phase, we will execute several key steps to prepare the CNN Daily Mail data for extractive text summarization using transformer models and reinforcement learning techniques. Initially, the extracted text from the dataset will undergo tokenization to segment it into meaningful units, facilitating further analysis. Sentence splitting will then be applied to divide the text into individual sentences, a necessary step for extractive summarization. Additionally, noise removal techniques will eliminate irrelevant elements, ensuring that only pertinent content is retained. Normalization processes will address issues such as capitalization, punctuation, and text variations. Optionally, common stop words may be removed, and data cleaning techniques will correct any encoding issues or inconsistencies in the text. To streamline the summarization process, we will define a maximum text length. These preprocessing steps are essential to ensure the quality and suitability of the CNN Daily Mail dataset for extractive text summarization using transformer models and reinforcement learning algorithms.

Tokenization

Tokenization is the process of breaking down a text into individual units, often words or sub-words. These units are called tokens. The goal is to convert a continuous text into discrete units that can be easily processed. For example, the sentence "The quick brown fox" would be tokenized into ("The", "quick", "brown", "fox").

Sentence Splitting

Sentence splitting involves dividing a continuous text into individual sentences. This step is crucial for tasks that require sentence level analysis. It helps in organizing and understanding the content at a more granular level

Noise Removal

Noise removal involves eliminating irrelevant characters, symbols, or elements from the text that do not contribute to the intended analysis. This may include removing special characters, HTML tags, or other nonessential elements depending on the context.

Normalization

Normalization aims to transform text into a standardized format. This can include converting all characters to lowercase, stemming (reducing words to their root form), and lemmatization (reducing words to their base or dictionary form). Normalization ensures consistency and reduces the complexity of the text.

Stop Word Removal

Stop words are common words (e.g., "the", "and" "is") that often do not carry significant meaning and can be removed without affecting the overall semantics of the text. Removing stop words can help reduce the dimensionality of the data and improve processing efficiency.

Data Cleaning

Data cleaning involves identifying and correcting errors or inconsistencies in the text data. This can include handling missing values, correcting typos, and addressing other issues that may arise during data collection or storage.

Text Length Limitation

Setting a limit on text length involves truncating or padding the text to a specific length. This step is often necessary when working with models that require fixed size inputs. It ensures that the input data is of consistent length, which is essential for efficient model training and inference.

MODEL SELECTION

This section combines two cutting edge approaches: BERT, a pre-trained transformer-based model, and Reinforcement Learning. BERT enables us to represent text in a highly context aware manner, and Reinforcement Learning fine tunes the summarization process based on a reward signal, ultimately leading to more precise and coherent extractive summaries. The integration of these two techniques has the potential to revolutionize the field of extractive text summarization by improving the quality of generated summaries.

BERT for Text Summarization

BERT (Bidirectional Encoder Representations from Transformers) is a pretrained deep learning model that captures contextual information from text. It processes text in both forward and backward directions, enabling it to understand the meaning of words in context. BERT's architecture includes multilayer transformers that create embeddings for each word in the input text, capturing complex relationships between words. BERT embeddings provide a strong foundation for extractive summarization, as they represent the context of words in a document, making it easier to identify important sentences for summarization.

RoBERTa (A Robustly Optimized BERT)

ROBERTA is a variant of BERT that was designed to address some of BERT's limitations and further optimize its pre-training process. It introduces modifications such as larger batch sizes, more training data, and longer training time. In the context of text summarization, RoBERTa's robust pre-training and improved generalization have proven beneficial. It excels in understanding context and is commonly used as a base model for various summarization tasks.

DistilBERT (Distilled BERT)

DistilBERT is a distilled and smaller version of BERT, created to be more efficient in terms of model size and computation. While it has fewer parameters than BERT, it retains much of the original model's performance. In summarization tasks, DistilBERT can provide a good balance between model size and accuracy, making it suitable for applications where computational resources are limited.

BART (Bidirectional and Auto Regressive Transformers)

BART is another transformer-based model that combines bidirectional and autoregressive pretraining. It has achieved success in abstractive text summarization, where it generates summaries from scratch. BART is not limited to extractive summarization and can be used for more creative summarization tasks.

PEGASUS

PEGASUS is a model designed specifically for abstractive text summarization. It is pretrained in a way that encourages generating coherent and fluent summaries from input text. PEGASUS has shown significant promise in generating humanlike abstractive summaries.

BERTSUM

BERTSUM is a framework for extractive summarization that combines the power of BERT with salience-based sentence scoring. It leverages BERT embeddings to represent sentences and applies a binary classification approach to determine which sentences are included in the summary.

Unified Transformers (UNIFIED)

UNIFIED is a unified model that combines both extractive and abstractive summarization capabilities. It uses transformer-based architectures and reinforcement learning for training. UNIFIED can be used for a wide range of summarization tasks, including single document and multi-document summarization.

Reinforcement Learning for Summarization

Reinforcement Learning is employed to fine tune the summarization process. In this framework, a reinforcement agent interacts with a model that generates summaries. The agent is trained to maximize a reward signal based on the quality of the generated summaries. It uses policy gradient methods or other reinforcement algorithms to adjust the model's behavior over time. Reinforcement Learning enables the model to learn how to select sentences that contribute to high quality summaries, as it receives feedback through the reward signal.

Combined Approach

The combined approach leverages BERT's contextual embeddings as features in the reinforcement learning framework. BERT embeddings are used to represent the source document and candidate summary sentences. The reinforcement agent interacts with these embeddings to make decisions about which sentences should be included in the summary. The use of BERT's context aware embeddings enhances the agent's ability to select sentences that are more relevant to the source text, leading to improved extractive summaries.

EVALUATION METRICS

The evaluation metrics for our extractive text summarization project utilizing a Transformer and Reinforcement Learning model include fundamental measures such as Rouge Score, Precision, Recall, and F1 Score, which assess the model's ability to accurately select and reproduce relevant information. Additionally, we will employ Cosine Similarity to gauge the similarity between generated summaries and reference summaries, and BLEU Score to measure the precision of the generated summary by comparing it to references. Semantic similarity metrics like Semantic Textual Similarity (STS) will provide insights into meaning closeness, while readability metrics such as Flesch-Kincaid Grade Level will assess the ease of comprehension. Coverage evaluation will ensure that essential information from the source document is adequately represented, and novelty metrics will gauge the inclusion of new information in the summary. Fluency will be assessed for coherence and grammatical correctness. Human evaluation through user feedback, including ratings and comments, will complement these metrics, providing a holistic understanding of the model's performance and practical utility.

CONCLUSION

In the ever-expanding realm of natural language processing, the pursuit of enhancing text summarization techniques has been relentless. This survey has explored a dynamic and promising avenue in the field: extractive text summarization, fortified by Transformer Based models like BERT and further optimized through Reinforcement Learning. The methodology section laid the groundwork by elucidating the profound significance of BERT and Reinforcement Learning in the summarization process. BERT, with its contextual understanding and fine-tuning capability, offers a solid foundation

for text representation, while Reinforcement Learning adds a layer of decision-making precision in extractive summarization. These methods bring text summarization closer to the elusive goal of automating the distillation of information from vast textual data, benefiting diverse domains such as news, academia, chatbots, and more.

We extend our gratitude to the pioneers, researchers, and innovators whose contributions have enriched this evolving field of knowledge. We eagerly look forward to the next chapter in the advancement of extractive text summarization.

REFERENCES

1. Alomari A, Idris N, Sabri AQ, Alsmadi I. Deep reinforcement and transfer learning for abstractive text summarization: A review. Comput Speech Lang. 2022 Jan 1; 71: 101276.
2. Wang G, Wu W. Surveying the landscape of text summarization with deep learning: A comprehensive review. Discrete Math Algorithms Appl. 2024;16(03):2330004. DOI: 10.1142/S1793830923300047.
3. Rawat A. Enhancing abstractive and extractive reviews text summarization using NLP and neural networks [Doctoral Dissertation]. Ireland: Dublin Business School; 2024.
4. Guan W, Smetannikov I, Tianxing M. Survey on automatic text summarization and transformer models applicability. In Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System. 2020 Oct 27; 176–184.
5. Gao Y, Meyer CM, Mesgar M, Gurevych I. Reward learning for efficient reinforcement learning in extractive document summarisation. arXiv preprint arXiv:1907.12894. 2019 Jul 30.
6. Ranganathan J, Abuka G. Text summarization using transformer model. In 2022 IEEE 9th International Conference on Social Networks Analysis, Management and Security (SNAMS). 2022, Nov; 1–5.
7. Srikanth A, Umasankar AS, Thanu S, Nirmala SJ. Extractive text summarization using dynamic clustering and coreference on BERT. In 2020 IEEE 5th international conference on computing, communication and security (ICCCS). 2020 Oct; 1–5.
8. Adhikari S. Nlp based machine learning approaches for text summarization. In 2020 IEEE Fourth International Conference on Computing Methodologies and Communication (ICCMC). 2020 Mar; 535–538.
9. Jugran S, Kumar A, Tyagi BS, Anand V. Extractive automatic text summarization using SpaCy in Python & NLP. In 2021 IEEE International conference on advanced computing and innovative technologies in engineering (ICACITE). 2021 Mar; 582–585.
10. Akhmetov I, Gelbukh A, Mussabayev R. Greedy optimization method for extractive summarization of scientific articles. IEEE Access. 2021; 9: 168141–168153.

