Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/101/20240946

Transformer models in text summarization

Xinrui Fu

School of Computing and Data science, Xiamen University Malaysia, Kuala Lumpur,43900, Malaysia

CST2109153@xmu.edu.my

Abstract. Text summarization represents a core research topic within the realm of natural language processing and is extensively applied across various domains, including journalism, library administration, information gathering, among others. With the development of deep learning, especially the proposed Transformer structure has greatly promoted the development of text summaries. This paper reviews the recent progress in Transformer-based text summarization methods. It begins with an overview of traditional text summarization techniques. The paper then delves into the advantages of Transformer models for text summarization, such as their ability to understand global context, dynamically allocate weights, and accelerate parallel computation. Text summarization models are classified into several types, such as abstraction-based, extraction-based, and those leveraging large language models. Notably, Models like PEGASUS, BERT, and HETFORMER have emerged as leading examples in this field. In addition, the effectiveness, advantages and disadvantages of these models are analyzed.

Keywords: Text summarization, transformer, deep learning.

1. Introduction

In an era of rapid information technology advancement, we find ourselves in a world with massive and continually growing volumes of information. In such a context, natural language processing (NLP) techniques, particularly text summarization, have become crucial tools to help us extract valuable information from vast amounts of text. The goal of text summarization is to automatically distill core content from lengthy text, providing users with a concise and accurate overview of information. Additionally, text summarization plays a crucial role in literature management, enabling researchers and scholars to efficiently extract key information from a vast corpus of documents. In the realm of information retrieval and search engines, text summarization enhances search experiences by providing users with quick previews of document content. In legal and medical domains, text summarization assists professionals in extracting essential information from complex documents, thereby improving work efficiency and ensuring accuracy, particularly in time-sensitive situations [1].

With the rapid advancement of deep learning technology, the field of text summarization has undergone revolutionary changes. Text summarization methods based on deep learning, particularly those utilizing the architecture of Transformer models, have made significant progress in extracting key information, generating coherent text, and understanding complex language structures. The core of these methods lies in simulating the human reading and comprehension process to capture the essence of the text more accurately. In recent times, numerous text summarization techniques have been introduced by

DOI: 10.54254/2755-2721/101/20240946

researchers, which are predominantly classified into two main categories: abstractive and extractive summarization. Abstractive summarization entails creating a summary by producing novel content that encapsulates the essence of the entire text. On the other hand, extractive summarization involves choosing key sentences directly from the text to form a representative summary.

Additionally, with the emergence of conversational large models (such as the GPT series), text summarization methods based on large pre-trained language models (LLMs) have also made significant progress [2]. These models, trained on large-scale text data, have learned rich language knowledge and contextual understanding.

Through a review of these advanced Transformer models and Large Language Models, this paper aims to provide readers with a comprehensive perspective to understand the latest advancements and potential applications of these models in the field of text summarization.

2. Preliminaries

2.1. Traditional Methods in Text Summarization

(1) **Graph Methods.** Inspired by PageRank, this method represents documents as graphs where sentences are vertices and similarity between them are edges. By partitioning the graph, it identifies topics and important sentences, useful for both single and multi-document summarization.

(2) **Latent Semantic Analysis (LSA).** LSA extracts semantic representations of text using singular value decomposition (SVD) on a matrix of words and sentences. It identifies topics within documents and is effective in both single and multi-document summarization, particularly in news.

(3) **Bayesian Topic Models.** These models represent document topics through probabilistic modeling, with LDA being the most advanced technique. LDA represents documents as mixtures of latent topics, aiding in identifying similarities and differences between documents.

(4) **ROUGE Evaluation Metrics.** ROUGE evaluates summary quality by comparing candidate and reference summaries using recall-based metrics like ROUGE-n, ROUGE-L, and ROUGE-SU, which focus on n-grams, longest common subsequence, and skip bigrams and unigrams allowing word insertion, respectively [3].

2.2. Transformer

The importance of the Transformer model in text summarization tasks is particularly prominent, primarily due to its efficient self-attention mechanism for capturing key textual information. In text summarization tasks, the goal is to extract the most important and core content from long texts, generating a concise overview. Traditional sequential models (such as RNNs) may be limited in their ability to model long-term dependencies, making it difficult to accurately focus on the key parts of the original text. However, the self-attention mechanism of the Transformer effectively addresses this issue:

(1) Global Context Understanding. The self-attention mechanism allows the model to inherently assess the significance of every term in the input sequence in relation to one another, rather than analyzing each word sequentially [4]. This is crucial for text summarization because key information may be distributed at any position in the text, and the model needs the ability to quickly locate and focus on this information.

(2) Dynamic Weight Allocation. Each word, when generating a summary, can receive different attention weights based on its importance in the entire context. This dynamic allocation of attention weights helps the model prioritize extracting the most representative and comprehensive information.

(3) Parallel Computation Acceleration. The Transformer model discards the recurrent structure and adopts a fully parallelized attention calculation method, greatly improving the speed of processing long texts, which is advantageous for handling the demands of large-scale text summarization.

Therefore, the Transformer model not only innovates the design of text summarization algorithms but also achieves higher-quality summary generation through its unique self-attention mechanism, promoting the development and practical application of text summarization technology. For instance, Transformer-based pre-trained models, including those from the BERT and GPT families, have shown

DOI: 10.54254/2755-2721/101/20240946

exceptional results on various text summarization benchmarks following suitable modifications and fine-tuning.

3. Text Summary Methods Based on Transformer

According to the type of text summary generation model, it can be divided into abstractive summary, extraction summary and based on large language model. The following describes the three types of models.

3.1. Abstractive Summarization

3.1.1. PEGASUS

The PEGASUS was proposed by Zhang et al [5].

The approach introduces an innovative self-supervised pre-training objective termed Gap Sentences Generation (GSG). In this method, key sentences are deleted or masked from the document, and the model must regenerate these sentences based on the context provided by the remaining content, similar to the process in extractive summarization. This technique is applied in the pre-training of large-scale Transformer-based encoder-decoder models, enhancing their capability to perform well on downstream summarization tasks.

(1) **Gap Sentences Generation.** GSG is the primary pre-training objective for the PEGASUS model. During this training, crucial sentences from the input document are obscured, and the model is challenged with the task of forecasting these concealed sentences. This technique mimics the extractive summarization process but goes a step further by producing fresh sequences of sentences rather than just replicating the text verbatim. GSG prompts the model to grasp the document's content as a whole and to produce summaries that are coherent, which is in line with the objectives of abstractive summarization.

(2) **Self-supervised learning.** Self-supervised learning is a technique for training models using data that doesn't need to be manually labeled. In PEGASUS, self-supervised learning is achieved through GSG. The model undergoes pre-training on extensive text corpora, and this process does not require manually annotated summary data. By guessing masked sentences, the model acquires a comprehension of generic language representations and structures. This knowledge is then applied to particular summarization tasks when the model is fine-tuned.

(3) **Multiple Importance Sentence Selection Strategies.** In the process of sentence generation, sentence masking can be selected in various ways, such as Random Selection, Lead Selection, Principal Selection, etc. These methods have different effects in different scenarios.

These three concepts collectively form the pre-training framework of the PEGASUS model, enabling it to master the ability to generate high-quality summaries even in the absence of abundant annotated summary data.

3.1.2. BERT-based method

The utilization of BERT in the realm of text summarization, especially within abstractive summarization, has shown remarkable progress, attributable to its groundbreaking architecture and effective fine-tuning techniques. BERT, as a pre-trained encoder, combined with a randomly initialized decoder, forms an encoder-decoder framework. This framework is capable not only of selecting key sentences from the source document but also of generating entirely new text sequences.

To more effectively combine the pre-trained encoder with the randomly initialized decoder, researchers have designed specialized fine-tuning strategies [6]. This includes using different optimizers and learning rate schedules to fine-tune the encoder and decoder separately. Furthermore, a two-stage fine-tuning process, which involves fine-tuning the encoder on an extractive summarization task first and then further refining it on an abstractive summarization task, can enhance the quality of the generated summaries.

DOI: 10.54254/2755-2721/101/20240946

Liu's proposed method for abstractive summarization is based on the pre-trained BERTSUM model and adopts an encoder-decoder architecture.[6]In this setup, BERTSUM acts as the encoder, tasked with deriving features from the input document, while the decoder, which consists of 6 layers of Transformer units, is employed to produce sentence vectors. To stabilize the training process, different learning rates are used to optimize the encoder and decoder. Specifically, the encoder uses a smaller learning rate to maintain the stability of the pre-trained parameters, while the decoder uses a larger learning rate to more effectively learn the generation task. This method leverages the pre-training advantages of BERTSUM while ensuring the stability of model training.This idea, which combines abstractive and extractive summarization, was proposed by Gehrmann et al. in 2018 [12], as evidenced by section 3.2.1.

In summary, BERT's application in text summarization, especially in abstractive summarization, has seen significant improvements due to its innovative structure and fine-tuning strategies. By employing an encoder-decoder framework and using different learning rates to optimize the encoder and decoder, BERT can effectively select key sentences from the source document and generate new text sequences. This approach makes full use of BERT's pre-training advantages while ensuring the stability of model training.

3.2. Extractive summarization

3.2.1. BERT-based method

In section 4.1.2, the application of the BERTSUM model in abstractive summarization was discussed. This section focuses on the application of BERT in extractive summarization [6].

The BERTSUM model demonstrates considerable potential in extractive summarization tasks. It utilizes BERT's document-level encoder to capture the semantic core of the document and incorporates [CLS] tokens and paragraph embeddings to differentiate sentences within the document, thereby gaining a deeper understanding of the contextual importance of each sentence. This enables the model to accurately extract key sentences. BERT's pre-training and fine-tuning paradigm, coupled with its bidirectional Transformer architecture, enhances the model's capacity to learn deep linguistic features, thereby improving the quality of summaries.

The BERTSUMEXT method proposed by Liu et al. further refines the model structure for extractive summarization, consisting of an encoder and a classifier. The encoder produces feature vectors for each sentence, while the classifier predicts the inclusion of sentences in the summary. The entire article is input into the model, and the encoder generates feature vectors for each sentence, which are then fed into a binary classifier for evaluation. The classifier's output determines the selection of sentences for the summary. The accuracy of distinguishing key sentences from non-essential ones is crucial for generating accurate and coherent summaries. The BERTSUM model and its variants have delivered exceptional performance in extractive text summarization tasks, providing significant support for automated summarization.

3.2.2. HETFORMER

The HETFORMER model incorporate Transformer's attention mechanism into traditional graph-based methods to enhance the performance of extractive summarization for long texts [7]. In earlier research on extractive summarization, graph-based methods such as TextRank would iteratively compute the significance of sentences by conceptualizing the document as a graph. In this graph, each node stands for a sentence, and the edges denote the similarity or connection between sentences [11]. This approach effectively identifies sentences that contribute most to the summary. The core innovations of the HETFORMER model include:

(1) **Heterogeneous graph representation**. HETFORMER treats words, sentences, and entities in the document as different types of nodes, constructing a heterogeneous graph. This representation allows the model to capture richer semantic information, as different types of nodes can represent different levels of semantic structure in the document.

DOI: 10.54254/2755-2721/101/20240946

(2) **Multi-granularity sparse attention**. Building upon Transformer, HETFORMER introduces multi-granularity sparse attention mechanisms, including token-to-token, token-to-sentence, sentence-to-sentence, and entity-to-entity attention patterns. These attention patterns enable the model to capture relationships between nodes at different granularities, thereby better understanding the structure and content of the document.

(3) **Transformer's self-attention**. HETFORMER utilizes Transformer's self-attention mechanism, allowing the model to consider not only local context information but also capture the global document structure when processing long texts. This combination of global and local information is crucial for generating high-quality summaries.

(4) **Pre-training and fine-tuning**. The HETFORMER model is pre-trained on an extensive collection of texts to learn universal language patterns. It is then further refined through fine-tuning on specific summarization challenges to enhance its effectiveness in these areas. This combination of pre-training and fine-tuning enables the model to better understand and utilize contextual information within the document.

Through these innovations, the HETFORMER model maintains the structured representation advantages of graph methods while enhancing its understanding and processing capabilities for long texts using Transformer's attention mechanism, thereby achieving significant performance improvements in long-text extractive summarization tasks.

3.2.3. HIBERT

This model also base on the BERT [8]. It focuses on improving extractive summarization through the following innovative aspects:

(1) **Hierarchical Bidirectional Encoder (HIBERT).** The document introduces a novel approach for pre-training hierarchical bidirectional Transformer encoders that is tailored for the representation of documents. This hierarchical encoder is used to obtain better sentence representations by considering the context of surrounding sentences.

(2) **Unsupervised Pre-training Objective.** The document introduces an unsupervised pre-training objective for HIBERT. It randomly masks out some sentences in a document and predicts the masked sentences using the context from other sentences in the document. This helps the model learn document-level representations.

(3) **Application to Extractive Summarization**. The pre-trained HIBERT encoder is fine-tuned for extractive summarization by adding a sentence classifier on top. The classifier predicts whether each sentence should be included in the summary or not.

(4) **Two-stage Pre-training**. The document utilizes two stages of pre-training - open-domain pre-training on a large dataset (GIGA-CM) and in-domain pre-training on the specific summarization dataset (CNNDM or NYT). Both stages are crucial for good performance.

5. State-of-the-art Results: The model suggested in the proposal delivers cutting-edge results on the CNN/Dailymail and New York Times summarization datasets, notably surpassing the performance of prior models.

In summary, the innovation lies in the development of HIBERT for document encoding and the unsupervised pre-training approach, which are specifically tailored for extractive summarization. The model does not combine generative and extractive methods.

3.3. Comparison of the effect of abstract generation methods

On the CNN/DailyMail dataset, the HETFORMERBase model's performance on the ROUGE-1, ROUGE-2, and ROUGE-L metrics is comparable to or surpasses that of current state-of-the-art baseline models. Specifically, HETFORMERBase achieves scores of 44.55, 20.82, and 40.37 on the ROUGE-1, ROUGE-2, and ROUGE-L metrics, respectively, demonstrating its competitiveness in single-document summarization tasks. On the Multi-News dataset, the HETFORMER model outperforms all extractive and abstractive baseline models on all three ROUGE metrics, indicating its effectiveness in avoiding significant information loss when handling longer documents.

DOI: 10.54254/2755-2721/101/20240946

Table 1. Comparison of effects between different Models

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
PEGASUSBASE (C4)	CNN/DailyMail	43.9	21.2	40.76
PEGASUSLARGE (C4)	CNN/DailyMail	44.17	21.47	41.11
HETFORMERBase	CNN/DailyMail	44.55	20.82	40.37
HETFORMERBase	Multi-News	46.21	17.49	42.43
HIBERT	New York Times	49.47	30.11	41.63
HIBERT	CNN/DailyMail	42.31	19.87	38.83
BERT	CNN/DailyMail	41.55	19.34	37.8
RoBERTa	CNN/DailyMail	42.99	20.6	39.21
BERTSUMEXT	CNN/DailyMail	43.25	20.24	39.63
BERTSUMABS	CNN/DailyMail	41.72	19.39	38.76
BERTSUMEXTABS	CNN/DailyMail	42.13	19.6	39.18
TRANS-ext	CNN/DM	41	18.4	36.9
TRANS-ext + filter	CNN/DM	42.8	21.1	38.4
TRANS-ext	Newsroom	37.2	25.2	32.4
TRANS-ext + filter	Newsroom	41.5	30.6	36.9

,

3.4. LLM in summarization

In the context of rapidly advancing natural language processing technologies, the article titled "Summarization is (Almost) Dead" suggests that conventional text summarization techniques may be nearing obsolescence due to the emergence of cutting-edge AI methods. The title implies that with the advent and application of large-scale pre-trained language models, traditional approaches to generating high-quality, coherent, and accurate summaries could be on the brink of becoming outdated. However, when referencing the actual content of the article, it's crucial to provide a nuanced interpretation, as the title might be provocative or overstated, while the body of the text would typically offer more comprehensive analysis and supporting evidence.

Large Language Models (LLMs), exemplified by GPT-3 variants and GPT-4, have been rigorously tested for their zero-shot summarization abilities across five distinct tasks. Datasets containing 50 novel samples per task were constructed to ensure no overlap with training data. In comparative evaluations involving human annotators, LLMs were assessed against fine-tuned models like BART, T5, and Pegasus [9] [10].

The study revealed that LLMs generally outperformed both human-written summaries and fine-tuned models in terms of overall summary quality. Pairwise comparisons and WinRateN M calculations highlighted LLMs' superiority. Fact consistency was also examined, showing that while LLMs like GPT-4 had commendable factual accuracy, human-written summaries occasionally presented factual inaccuracies, especially in complex tasks like multi-news and code summarization.

4. Conclusion

Transformer-based text summarization models, such as PEGASUS and BERT-based approaches, represent the cutting-edge technology in the field. PEGASUS is capable of generating rich and high-quality summaries through its unique Gap Sentences Generation pre-training objective. BERT-based methods leverage an encoder-decoder framework and incorporate the pre-training advantages of the BERT model to further enhance the accuracy and fluency of summaries. Additionally, the HETFORMER model, as a hierarchical dual-attention model, is particularly suitable for long text

DOI: 10.54254/2755-2721/101/20240946

summarization tasks, effectively overcoming the limitations of traditional models in handling long documents.

With the emergence of large language models like GPT, text summarization methods based on large pre-trained language models have made significant progress. These models, pre-trained on large-scale text data, possess deep language understanding and context-awareness capabilities. Experimental data shows that the HETFORMER model performs exceptionally well on the New York Times dataset, and other models also successfully complete text summarization tasks.

Although current evaluation methods such as ROUGE play an important role in measuring the similarity between generated summaries and reference summaries, these metrics are insufficient in assessing the semantic correctness and credibility of summaries. Therefore, future research should focus on developing more comprehensive evaluation methods and credibility detection mechanisms to ensure the accuracy and authenticity of summaries. This will contribute to the construction of a more precise and reliable evaluation system for text summarization, providing clear directions for research in the field.

References

[1] Allahyari M, Pouriyeh S, Assefi M, et al(2017). Text Summarization Techniques: A Brief Survey. International Journal of Advanced Computer Science and Applications, 8(10).

[2] Phillips T, Saleh A, Glazewski K D, et al. Exploring the use of GPT-3 as a tool for evaluating text-based collaborative discourse. Companion Proceedings of the 12th, 2022, 54.

[3] El-Kassas W S, Salama C R, Rafea A A, et al (2021). Automatic text summarization: A comprehensive survey. Expert systems with applications, 165: 113679.

[4] Vaswani A, Shazeer N, Parmar N, et al(2017). Attention is all you need. Advances in neural information processing systems, 30.

[5] Zhang J, Zhao Y, Saleh M A, et al(2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. International Conference on Machine Learning, 1: 11328–11339.

[6] Liu Y, Lapata M(2021). Text Summarization with Pretrained Encoders. arXiv preprint arXiv:1908.08345, 2019.

[7] Liu Y, Zhang J G, Wan Y, et al(2019). HETFORMER: Heterogeneous transformer with sparse attention for long-text extractive summarization. arXiv preprint arXiv:2110.06388.

[8] Zhang X, Wei F, Zhou M. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. arXiv preprint arXiv:1905.06566.

[9] Pu X, Gao M, Wan X(2023). Summarization is (Almost) Dead. arXiv preprint arXiv:2309.09558.

[10] Zhang T, Ladhak F, Durmus E, et al(2020). Benchmarking Large Language Models for News Summarization. Transactions of the Association for Computational Linguistics, 2024, 12: 39-57.

[11] Jin H, Wang T, Wan X. Multi-granularity interaction network for extractive and abstractive multi-document summarization.//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 6244-6254.

[12] Gehrmann S, Deng Y, Rush A M(2018). Bottom-up abstractive summarization. arXiv preprint arXiv:1808.10792.