IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305073 |

A Comprehensive Overview of Pre-trained Models for Text Summarization

Soham Anand Karulkar

Student, Department of Electronics and Computer Science, Vidyalankar Institute of Technology, Mumbai, India

ABSTRACT: This paper provides a comprehensive overview of automatic text summarization, a critical task in natural language processing aimed at condensing large volumes of text into concise summaries. The document discusses two primary approaches to text summarization, extractive and abstractive, and explores the role of pre-trained models in enhancing summarization performance. Various pre-trained models, including BART, BERT, DistilBART, GPT-2, T5, PEGASUS, GPT-3, PALM, and RoBERTa, are examined, highlighting their architectures, capabilities, and applications in text summarization tasks. The paper also addresses the challenges of information overload in the digital age and underscores the importance of automated summarization tools in managing vast amounts of textual data. Overall, this paper serves as a valuable resource for researchers and practitioners interested in the latest advancements and trends in automatic text summarization.

KEYWORDS: Automatic Text Summarization, Extractive Summarization, Abstractive Summarization, Pre-Trained Models, BART, BERT, DistilBART, GPT-2, T5, PEGASUS, GPT-3, PALM, RoBERTa, Natural Language Processing, Information Overload, Digital Age, NLP Applications.

I. INTRODUCTION

In the era of information abundance facilitated by the internet, the task of distilling vast amounts of text into concise, meaningful summaries has become increasingly vital. Automatic text summarization, a key component of natural language processing (NLP), offers a solution to this challenge by condensing lengthy articles, documents, or news stories into succinct summaries, thereby facilitating quick information retrieval and mitigating the effects of information overload. This paper explores the foundational concepts of text summarization, focusing on two primary approaches: extractive summarization, which selects and combines key sentences from the source text, and abstractive summarization, which generates summaries by understanding the content and context of the input text. Additionally, the role of pre-trained models in text summarization is examined, highlighting their ability to learn and generalize from large text corpora, enabling them to generate coherent and informative summaries with high efficiency and accuracy. The paper also provides an overview of several prominent pre-trained models in NLP, such as BART, BERT, and GPT-3, discussing their architectures, functionalities, and contributions to the field of automatic summarization. Through this exploration, the paper aims to elucidate the advancements and challenges in text summarization, offering insights into the evolving landscape of NLP and its applications in managing and extracting knowledge from vast textual data sources.[1]

II. FOUNDATIONS OF TEXT SUMMARIZATION AND THE ROLE OF PRE-TRAINED MODELS

A. What is text summarization ?

Text summarization is a process in natural language processing (NLP) where a machine learning model is used to condense a piece of text, such as an article, document, or news story, into a shorter version while still retaining the key information and main points of the original text. This condensed version, known as a summary, provides a quick and concise overview of the text, making it easier for readers to understand the main ideas without having to read the entire document.[2]

B. What are pre-trained models?

Pre-trained models are machine learning models that have been trained on a large corpus of text data to perform a specific NLP task, such as text summarization, without the need for additional training on the specific dataset. These models have learned the underlying patterns and structures of language from the training data and can be fine-tuned on a smaller dataset for a specific task to improve their performance.[3]

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305073 |

C. How pre-trained models can be utilized for text summarization?

Pretrained models can be used for text summarization by fine-tuning them on a specific dataset related to the summarization task (if required). This involves taking a pre-trained model and updating its parameters using a smaller dataset of summarization examples. During the fine-tuning process, the model learns the specifics of the summarization task, such as identifying key information and generating concise summaries based on the input text. Fine-tuned models can then be used to automatically generate summaries for new pieces of text, saving time and effort compared to manual summarization methods.

III. PRETRAINED MODELS

D. BART (Bidirectional and Auto-Regressive Transformers)

BART, an innovative model in the field of natural language processing (NLP), was introduced by Facebook AI researchers in 2019. It quickly garnered attention for its ability to excel in various NLP tasks, including text summarization, machine translation, and text generation. Built upon the Transformer architecture, BART leverages powerful attention mechanisms to capture long-range dependencies in input sequences. Unlike traditional models, BART focuses on sequence-to-sequence tasks, requiring both an encoder to process input data and a decoder to generate output. However, for tasks like text summarization, BART utilizes only the encoder component. The model's effectiveness lies in its ability to learn rich representations of input text, enabling it to generate coherent and informative summaries. Fine-tuning BART on specific datasets further enhances its performance, making it a versatile and valuable tool in NLP applications.[4]

E. BERT (Bidirectional Encoder Representations from Transformers)

BERT, a landmark in NLP, was introduced by Google AI Language researchers in October 2018, [5]sparking significant interest in the field. It achieved state-of-the-art performance across various NLP tasks, including Question-Answering, Sentiment Analysis, and language translation. As a Transformer model, BERT utilizes attention mechanisms to understand the relationships between words in a sequence. Unlike traditional models, Transformers consist of two key components: an encoder, which processes input data, and a decoder, which generates output. Since BERT focuses on language modelling, it only requires the encoder. The specific workings of Transformers are detailed in the original research paper.[6]

F. DistilBART

DistilBART, a variant of the BART model, was introduced as a more lightweight and efficient version by Hugging Face in 2020. This model aims to retain the key capabilities of BART while reducing its size and computational requirements, making it more accessible for a wider range of applications. Like BART, DistilBART is based on the Transformer architecture and is designed for sequence-to-sequence tasks, such as text summarization and machine translation. It consists of an encoder-decoder structure, with the encoder responsible for processing input data and the decoder for generating output. DistilBART's compact size and efficient training make it a practical choice for tasks where computational resources are limited. Despite its smaller size, DistilBART maintains competitive performance compared to its larger counterparts, making it a valuable option for various NLP applications.[7]

G. GPT-2 (Generative Pre-trained Transformer 2)

GPT-2, short for "Generative Pre-trained Transformer 2," was introduced by OpenAI in 2019 and represents a significant advancement in natural language processing (NLP). This model gained attention for its ability to generate coherent and contextually relevant text, showcasing its effectiveness in various NLP tasks. Built on the Transformer architecture, GPT-2 is a large-scale model trained on a diverse dataset of text from the internet. It employs a decoder-only architecture, where the model processes input sequentially and generates output one token at a time. GPT-2's key innovation lies in its ability to generate human-like text by predicting the next word in a sequence based on the context provided by the preceding words. This approach allows GPT-2 to excel in tasks such as text generation, language translation, and question answering. Despite its impressive performance, GPT-2 has raised concerns about the potential misuse of AI-generated text, leading to responsible deployment practices in its usage.[8], [9]

H. T5

T5, or Text-to-Text Transfer Transformer, is a versatile and powerful model introduced by Google AI in 2019. T5 differs from traditional models in that it frames all NLP tasks as text-to-text tasks, where both the input and output are textual. This approach allows T5 to handle a wide range of tasks, including text summarization, machine translation, and question answering, in a unified manner. T5 is based on the Transformer architecture and consists of an encoder-decoder structure, similar to models like BERT and GPT-2. However, T5's innovative text-to-text framework enables it to

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305073 |

achieve state-of-the-art performance on many NLP benchmarks. By training on a diverse set of text-to-text tasks, T5 learns to generalize across tasks and domains, making it a highly effective and adaptable model for various NLP applications.[10]

I.PEGASUS

PEGASUS, introduced by Google Research in 2020, is a transformer-based model designed specifically for abstractive text summarization. PEGASUS builds upon the BERT architecture but introduces several key innovations to improve its performance in summarization tasks. One of the key features of PEGASUS is its novel pre-training approach, which involves denoising documents instead of individual sentences or spans of text. This allows PEGASUS to generate more coherent and informative summaries by understanding the context of the entire document. Additionally, PEGASUS utilizes a self-supervised objective called gap-sentence generation, which helps the model learn to generate summaries by predicting missing sentences in a document. Overall, PEGASUS has achieved state-of-the-art performance on various text summarization benchmarks, demonstrating its effectiveness in generating high-quality summaries.[11]

J. GPT 3

GPT-3, the successor to GPT-2, was introduced by OpenAI in 2020. It is known for its massive scale, with 175 billion parameters, making it one of the largest language models ever created. GPT-3 further improves upon its predecessors by demonstrating superior performance on a wide range of NLP tasks, including text summarization. Like GPT-2, GPT-3 is based on the Transformer architecture and utilizes unsupervised learning. However, GPT-3's larger size allows it to generate even more coherent and contextually relevant text, often indistinguishable from human-written text. Despite its impressive capabilities, GPT-3 also raised concerns about potential misuse and ethical implications due to its ability to generate highly convincing fake text.[12], [13]

K. PALM

PALM, short for Pre-trained Autoencoding Latent Model, was introduced by the Hong Kong University of Science and Technology (HKUST) and Peking University in 2020. PALM is designed for text generation tasks, including text summarization, language modeling, and dialogue generation. It is based on the Transformer architecture, which allows it to model long-range dependencies in text efficiently What sets PALM apart is its use of an autoencoding framework, where the model is trained to reconstruct the input text from a corrupted version of itself. PALM can be used for text summarization by leveraging its ability to generate coherent and informative summaries. During training, PALM learns to reconstruct input text from a corrupted version of itself, capturing the key information and context. This learned ability allows PALM to generate concise summaries by selecting the most relevant information from the input text.[14]

L. Excellent RoBERTa

Excellent Roberta is a variant of the RoBERTa model, which stands for Robustly optimized BERT approach. RoBERTa, introduced by Facebook AI in 2019, builds upon the BERT model by optimizing key hyperparameters, training it on more data, and for longer durations. This leads to improved performance on various NLP tasks, including text summarization, by better capturing nuanced language patterns. RoBERTa's enhancements enable it to achieve state-of-the-art results on benchmarks like GLUE and SQuAD, demonstrating its effectiveness in understanding and generating human language.[15]

IV. CONCLUSION

The rapid expansion of the Internet has led to an overwhelming volume of available information, posing challenges for human summarization. Consequently, there is a pressing need for automated summarization tools to address the information overload. This paper focuses on extractive methods for summarizing single and multiple documents, highlighting widely-used techniques such as topic representation, frequency-driven approaches, graph-based algorithms, and machine learning methods. While it is not possible to comprehensively cover all diverse algorithms and approaches in this paper, it aims to offer valuable insights into recent trends and advancements in automatic summarization, presenting a snapshot of the state-of-the-art in this field of research.

REFERENCES

[1] J. Li, C. Zhang, X. Chen, Y. Cao, P. Liao, and P. Zhang, Abstractive Text Summarization with Multi-Head Attention. [Online]. Available: http://www.ieee.org/publications

[2] D. Huang et al., “What Have We Achieved on Text Summarization?,” Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.04529

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305073 |

[3] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained Models for Natural Language Processing: A Survey," Mar. 2020, doi: 10.1007/s11431-020-1647-3.
[4] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.13461
[5] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor
[6] S. Singla, "Comparative Analysis of Transformer Based Pre-Trained NLP Models Article in," International Journal of Computer Sciences and Engineering Open Access Research Paper, vol. 8, no. 11, 2020, doi: 10.26438/ijcse/v8i11.4044.
[7] S. Alaei, "An Automated Discharge Summary System Built for Multiple Clinical English Texts by Pre-trained DistilBART Model."
[8] P. Budzianowski and I. Vulić, "Hello, It's GPT-2 -- How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.05774
[9] V. Kieuvongngam, B. Tan, and Y. Niu, "Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.01997
[10] M. Borah, P. Dadure, and P. Pakray, "Comparative analysis of T5 model for abstractive text summarization on different datasets." [Online]. Available: https://ssrn.com/abstract=4096413
[11] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," 2020.
[12] S. Kublik and S. Saboo, Gpt-3 : The Ultimate Guide to Building NLP Products with OpenAI API.
[13] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," Minds and Machines, vol. 30, no. 4. Springer Science and Business Media B.V., pp. 681–694, Dec. 01, 2020. doi: 10.1007/s11023-020-09548-1.
[14] B. Bi et al., "PALM: Pre-training an Autoencoding&Autoregressive Language Model for Context-conditioned Generation," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.07159
[15] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," Applied Intelligence, vol. 51, no. 6, pp. 3522–3533, Jun. 2021, doi: 10.1007/s10489-020-01964-1.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details