Knowledge Extraction: Pegasus-driven Summarization for Research Dissemination

Asha Hegde[1], Sharal Coelho[1] and Joiesmary D'Souza[2]

[1]Department of Computer Science, Mangalore University, India
[2]Manipal Academy of Higher Education, Manipal, India, India

Abstract

Many research articles in recent years have included research highlights to concisely summarize important findings, making it easier for academics to understand a paper's contribution and increasing the article's visibility through search engines. The field of automatic text processing research using Artificial Intelligence is very active recently. The proposed work uses a refined Pegasus model with Low-Rank Adaptation (LoRA) to summarize scientific abstracts into highlights for effective and high-quality summary generation. The process involves pre-processing and tokenizing the MixSub-SciHigh dataset provided by "Research Highlight Generation from Scientific Papers (SciHigh)" shared task, training with model variants from ArXiv and PubMed to provide accurate highlights from various scientific domain. Among the two models, the proposed summarization model that uses PubMed with LoRA obtained a ROUGE-L F1-score of 0.2296 and secured 4th rank in the shared task.

Keywords

Text Summarization, MixSub, Research Highlight Generation, Pegasus/PubMed

1. Introduction

The scientific research process begins with investigating the state of the art, which may involve a vast number of publications [1]. However, the growth in the number of scientific papers is generally viewed as a positive development, but it also shows some challenges. With the exponential growth in the number of scientific papers being published, it can be challenging for researchers to stay up to date with the latest findings in their field and identify which papers are most important.

Summarization is the process of reducing a source text into a shorter version while preserving its information content [2]. Automatically summarizing scientific articles helps researchers in their investigation by speeding up the research process [3]. However, generating summaries of scientific articles addresses the problems, such as (i) identifying the keywords that describe the main topics covered by an article, (ii) generating an abstract of an article, or (iii) selecting the content that is most likely to appear in the article highlights. Keywords are single words or phrases, which are usually extracted using keyphrase extraction techniques. Abstracts are collections of whole sentences that provide summaries of the most important information in a publication. Although they are commonly available for most of the published articles, they can also be generated using extractive summarization methods [4].

Research highlights are straightforward, structured summaries that encapsulate the essential contributions of a scientific paper, offering a quick overview for students, journals, and researchers. Manually preparing highlights is time-consuming, prompting the need for automated solutions. The shared task Research Highlight Generation from Scientific Papers (SciHigh) focuses on automatically generating research highlights from scientific paper abstracts. This shared task aims to develop Machine Learning (ML) models to generate high-quality, author-like research highlights from abstracts using the MixSub dataset [5], a subset of the MixSub corpus comprising 19,785 scientific papers from 2020. We explore transformer-based models, retrieval-augmented approaches, and fine-tuned neural networks to achieve this goal, addressing challenges such as incorrect information and factual inconsistencies. The task evaluates submissions using ROUGE-1, ROUGE-2, ROUGE-L, and METEOR metrics, with rankings

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
hegdekasha@gmail.com (A. Hegde); sharalmucs@gmail.com (S. Coelho); joiesdsouza@gmail.com (J. D'Souza)

based on the ROUGE-L F1-score. Our work aims to improve the efficiency and accuracy of highlight generation, benefiting academic workflows and knowledge dissemination.

The rest of the paper is organized as follows: Section 2 contains Related Work. While Section 3 describes the Methodology, Section 4 gives a description of the Datasets, Experiments, Results, and Observations followed by Conclusion in Section 5.

2. Related Work

Automatic text summarization has been extensively studied[6], with significant advancements in abstractive summarization driven by transformer-based models [7][8]. Models like BART [9], T5 [10], and PEGASUS [11] have shown strong performance in generating summaries for general and scientific texts. While both BART and PEGASUS are designed for abstractive tasks, PEGASUS is better suited for highlight-like outputs because it has been pre-trained on gap-sentence generation. Recent work on scientific summarization, such as SciBERT [12] and PubMedBERT [13], Utilizes domain-specific pre-training to effectively manage specialized terminology, which is critical for the MixSub-SciHigh dataset's diverse scientific domains.

Retrieval Augmented Generation (RAG) [14] combines retrieval and generation to improve accurate consistency, as seen in studies like Multi-XScience [15] and SciTLDR[16]. RAG retrieves relevant context to ground outputs, addressing issues like hallucinations. Fine-tuning neural networks on domain-specific datasets has also proven effective, with studies demonstrating improved ROUGE scores through targeted fine-tuning on scientific texts. However, generating structured highlights remains underexplored, and challenges like factual inconsistencies persist in scientific summarization. Nallapati et al. [17] proposed an abstractive text summarization technique that uses attentional encoder-decoder Recurrent Neural Networks (RNN). This model is used to generate a summary of a given input document. Using a bidirectional RNN, the model first encodes the input document, which captures the input's contextual information. At the decoder end, the summary is then generated one word at a time by considering the encoded input document. Our study builds on these advancements by focusing on highlight generation, a summarization task, using the MixSub-SciHigh dataset. We aim to combine transformer-based models, retrieval-augmented approaches, and fine-tuning strategies to optimize for ROUGE-L F1-score.

3. Methodology

The objective of this work to generate research highlights using the abstraction of the research articles. The methodology involves fine-tuning a pre-trained Pegasus model using Low-Rank Adaptation (LoRA) for abstractive summarization of scientific articles, where abstracts are summarized into highlights. The process begins with preprocessing steps followed by model building. Initially, the Pegasus-PubMed model is used, and subsequently, it is replaced with the Pegasus-ArXiv model to compare performance on domain-specific summarization tasks.

3.1. Pre-processing

The dataset consists of columns 'Abstract' and 'Highlights' which represent the input and target values, respectively. The Pegasus tokenizer is applied to both inputs and targets, truncating to maximum lengths of 256 and 64 tokens, respectively, with padding to ensure uniform sequence lengths. This tokenized data is then mapped across the datasets, removing original columns to create model-ready inputs.

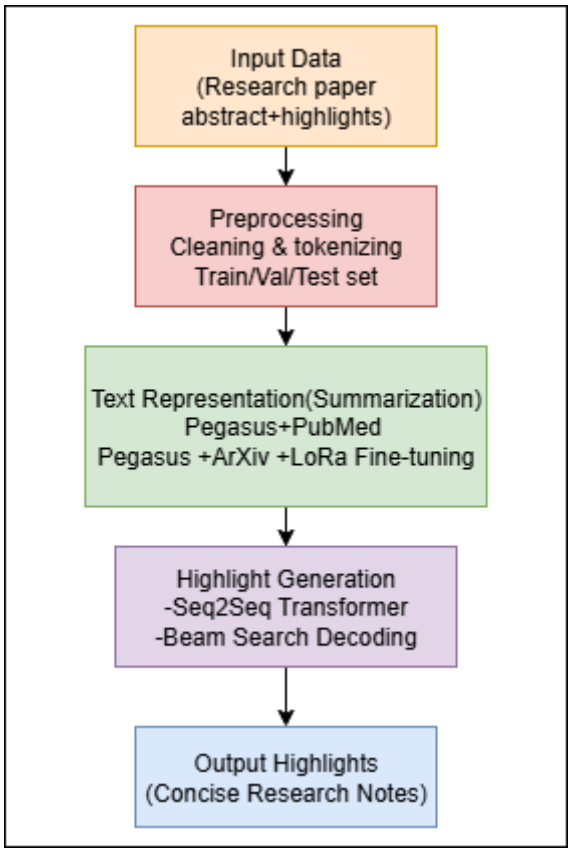

Figure 1: Block diagram of the proposed method

3.2. Text Representation Using Pegasus/PubMed

Pegasus[1] is a pre-trained transformer-based model designed specifically for abstractive text summarization [18]. It uses a sequence-to-sequence architecture where the encoder processes the input text and the decoder generates the summary, usig self-supervised pre-training on large corpora with a gap-sentence generation objective to mimic summarization tasks. Its pre-training focuses on removing important sentences from documents and generating them as pseudo-summaries, which improves its ability to capture important points. The significance of the PubMed variant lies in its fine-tuning on the PubMed dataset, a vast collection of biomedical literature abstracts, which provides it with domain-specific knowledge in medical and life sciences terminology.

3.3. Text Representation Using Pegasus/ArXiv

Pegasus is used due to its efficiency in abstractive summarization tasks. The ArXiv variant is fine-tuned on the ArXiv dataset, which comprises millions of preprints in fields like physics, mathematics, computer science, etc., allowing the model to specialize in non-biomedical scientific literature.

3.4. LoRA

LoRA[19] is a Parameter-efficient Fine-tuning (PEFT) method that maintains competitive performance while drastically reducing memory and computing expenses. In summarization tasks, LoRA's significance lies in its ability to fine-tune large models like Pegasus on domain-specific data with limited computational resources. This is particularly beneficial for scientific summarization, where datasets may be smaller or specialized, as LoRA preserves the model's generalization while adapting to nuances

[1]https://huggingface.co/docs/transformers/model_doc/pegasus

Table 1
Distributions of the MixSub-SciHigh dataset

Training Set	Validation Set	Test Set
10,000	1985	1840

like technical terminology. Furthermore, it facilitates experimentation, such as switching between PubMed and ArXiv variants, by making the process more efficient and scalable. Overall, LoRA helps in reducing memory footprint and training time without sacrificing summary quality.

3.5. Model Construction

The model is constructed by loading the pre-trained Pegasus base from either the PubMed or ArXiv checkpoint. LoRA configuration is applied with r=8, alpha =16, and dropout =0.1. A Seq2SeqTrainer is initialized with the adapted model. The training arguments includes batch sizes, learning rate, and epochs, along with a data collator.

4. Experiments an Result

4.1. Dataset

The proposed work utilizes a subset of the MixSub dataset. The MixSub corpus is created by collecting research articles from ScienceDirect[2], encompassing a diverse range of scientific domains. It contains 19,785 research papers published in the year 2020. The MixSub-SciHigh dataset is split into three sets and detailed distribution is shown in table 1. Each data instance is structured as a pair consisting of the abstract and the corresponding author-written research highlights. Each entry in the dataset includes:

- **Abstract**: A concise summary of the research paper.
- **Research Highlights**: Key contributions manually written by the authors.

4.2. Experimental Setup

The experiments are performed using the MixSub-SciHigh dataset from the "Research Highlight Generation from Scientific Papers (SciHigh)" shared task. The dataset comprises scientific abstracts from diverse fields, pre-processed and tokenized for model training. Two model variants are evaluated: fine-tuned on (i) ArXiv data and (ii) PubMed data, both utilizing the Pegasus model with LoRA for efficient abstract summarization. The performance of the models is evaluated using ROUGE metrics such as ROUGE-1, ROUGE-2, and ROUGE-L to measure the quality and relevance of the generated research highlights.

4.3. Results

The performance of the two model variants is summarized in the Table 2, based on ROUGE F1-scores for the generated highlights. Figure 2 presents a bar chart comparing our team's (NLPFusion) best submission against other teams' submissions in the shared task, highlighting our competitive performance in ranking quality.

The performance of our 2 submission runs for the SciHigh task is presented in Table 2. The PubMed-based model outperformed the ArXiv-based model across all ROUGE metrics, achieving a ROUGE-L F1-score of 0.2296, which secured the 4th rank in the SciHigh shared task. The higher performance of the PubMed model can be attributed to its fine-tuning on biomedical literature, which aligns closely with the

[2]https://www.sciencedirect.com/

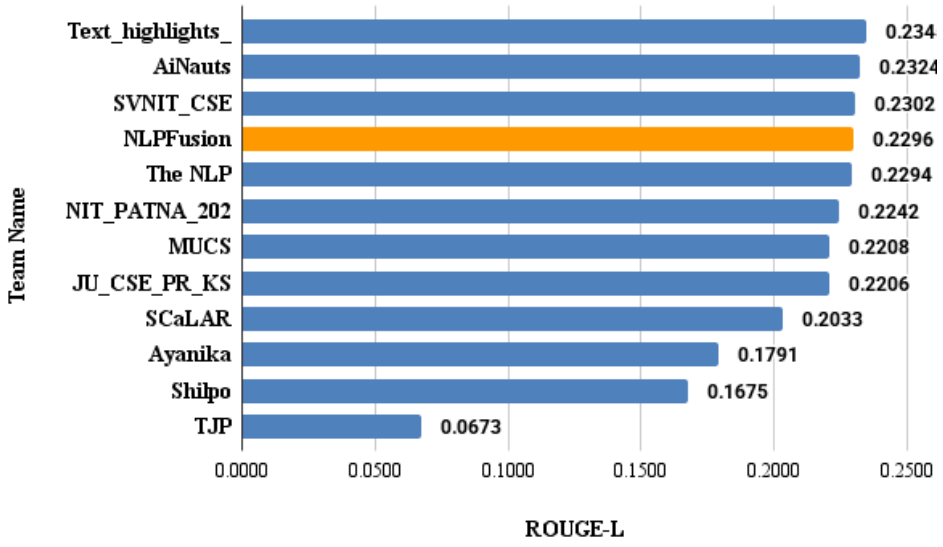

Figure 2: Comparison of Performance Across Team Submissions

Table 2
Performance of the proposed models

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Rank
pegasus/arxiv+LoRA	0.2885	0.0917	0.2040	0.2210	-
pegasus/pubmed+LoRA	0.3214	0.1168	**0.2296**	0.2926	**4**

domain-specific characteristics of the MixSub-SciHigh dataset. The results demonstrate the effectiveness of the LoRA-based fine-tuning approach in generating accurate research highlights, contributing to efficient summarization of scientific abstracts.

5. Conclusion

The task of automatically generating research highlights from abstracts represents an important step toward facilitating academic workflows. By using transformer-based models, our team, NLPFusion, aims to produce high-quality, author-like highlights that achieve strong ROUGE-L F1-scores while addressing challenges like hallucinations and factual inconsistencies. The given dataset's diverse scientific domains and structured highlight format provide a robust research facility for advancing summarization techniques. Our work has the potential to reduce the burden on researchers and enhance the accessibility of scientific contributions through automated, accurate highlight generation.

Declaration on Generative AI

In preparing this work, the author(s) utilized Grok[3] for grammar and spelling checks. Paraphrasing was handled via QuillBot. With this tool, the author(s) reviewed and revised the content as required, while assuming full responsibility for the publication's integrity.

[3]https://grok.com

References

[1] T. Rehman, D. K. Sanyal, S. Chattopadhyay, Research highlight generation with elmo contextual embeddings, Scalable Computing: Practice and Experience 24 (2023) 181–190.

[2] K. Woodsend, M. Lapata, Automatic generation of story highlights, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 565–574.

[3] T. Rehman, S. Das, D. K. Sanyal, S. Chattopadhyay, An analysis of abstractive text summarization using pre-trained models, in: Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2021, Springer, 2022, pp. 253–264.

[4] S. Chopra, M. Auli, A. M. Rush, Abstractive sentence summarization with attentive recurrent neural networks, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 93–98.

[5] T. Rehman, D. K. Sanyal, S. Chattopadhyay, P. K. Bhowmick, P. P. Das, Generation of highlights from research papers using pointer-generator networks and scibert embeddings, IEEE Access 11 (2023) 91358–91374.

[6] T. Rehman, S. Chattopadhyay, D. K. Sanyal, Abstractive summarization of scientific documents: Models and evaluation techniques, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023, pp. 121–124.

[7] A. S. Bashir, A. A. Bichi, U. Mahmud, A. M. Bello, Long-text abstractive summarization using transformer models: A systematic review, Journal of the Brazilian Computer Society 31 (2025) 1264–1279.

[8] J. Jons, Text summarization using a transformer architecture: An attention based transformer approach to abstractive summarization, 2024.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.

[11] J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International conference on machine learning, PMLR, 2020, pp. 11328–11339.

[12] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[13] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023).

[15] Y. Lu, Y. Dong, L. Charlin, Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles, arXiv preprint arXiv:2010.14235 (2020).

[16] I. Cachola, K. Lo, A. Cohan, D. S. Weld, Tldr: Extreme summarization of scientific documents, arXiv preprint arXiv:2004.15011 (2020).

[17] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al., Abstractive text summarization using sequence-to-sequence rnns and beyond, arXiv preprint arXiv:1602.06023 (2016).

[18] N. V. K. S. Dasari, A. Sungheetha, R. Sharma, G. L. V. Mahesh, G. Danesh, T. Singh, Text summarization using pegasus transformer model in machine learning, in: 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), IEEE, 2025, pp. 786–788.

[19] J. Y.-C. Hu, M. Su, E.-J. Kuo, Z. Song, H. Liu, Computational limits of low-rank adaptation (lora) for transformer-based models, arXiv preprint arXiv:2406.03136 (2024).