Two-Stage Graph-Augmented Summarization of Scientific Documents

Rezvaneh Rezapour[1], **Yubin Ge**[2], **Kanyao Han**[2], **Ray Jeong**[2], **Jana Diesner**[2,3]
[1] Drexel University
[2] University of Illinois at Urbana-Champaign
[3] Technical University of Munich
shadi.rezapour@drexel.edu jana.diesner@tum.de
{yubinge2, kanyaoh2, yj25}@illinois.edu

Abstract

Automatic text summarization helps to digest the vast and ever-growing amount of scientific publications. While transformer-based solutions like BERT and SciBERT have advanced scientific summarization, lengthy documents pose a challenge due to the token limits of these models. To address this issue, we introduce and evaluate a two-stage model that combines an extract-then-compress framework. Our model incorporates a "graph-augmented extraction module" to select order-based salient sentences and an "abstractive compression module" to generate concise summaries. Additionally, we introduce the *BioConSumm* dataset, which focuses on biodiversity conservation, to support underrepresented domains and explore domain-specific summarization strategies. Out of the tested models, our model achieves the highest ROUGE-2 and ROUGE-L scores on our newly created dataset (*BioConSumm*) and on the *SUMPUBMED* dataset, which serves as a benchmark in the field of biomedicine.

1 Introduction

The scientific community has experienced an unprecedented surge in the number of scientific publications (Erera et al., 2019). This exponential growth has resulted in a state of information overload, presenting both opportunities and challenges for researchers. Navigating the vast amount of information, filtering out relevant studies, and extracting essential insights have become increasingly challenging (Levy, 2008). To address this issue, researchers have turned to automatic summarization methods, which leverage various NLP techniques to condense the essential findings, methodologies, and contributions of research articles into concise and informative summaries.

The use of transformer-based language models (LMs), such as BERT (Devlin et al., 2019), BART (Lewis et al., 2020), SciBERT (Beltagy et al., 2019), and T5 (Raffel et al., 2020) has significantly improved text summarization. Particularly SciBERT's performance in handling science-related documents, and BERTSUM (Zhong et al., 2020), an extension of BERT for extractive summarization, have advanced scientific summarization in terms of domain-specific performance, accuracy, and coherency (Liu, 2019; Sefid and Giles, 2022). Large language models (LLMs) have further reshaped the field of text summarization. For example, OpenAI's GPT-3 and its successors (Achiam et al., 2023) have shown remarkable capabilities in generating coherent and informative summaries (Tang et al., 2023; Jin et al., 2024). These models can perform both extractive and abstractive summarization with high accuracy and fluency. For instance, Zhang et al. (2024) highlight how LLMs can generate summaries that not only capture the main ideas of the source text but also reformulate them in novel ways, often providing additional context or explanations.

Despite recent advancements in text summarization, there are strong reasons to continue refining BERT-based models for scientific summarization: SciBERT, for instance, is trained on scientific texts, which provides with an edge in identifying academic papers' unique language and structure. They are also computationally more efficient and have smaller memory requirements than LLMs, making them faster and more accessible for processing large volumes of scientific papers (Zhu et al., 2023). Additionally, such models offer greater interpretability, allowing for insights into the summarization process, which is crucial in the context of science (White et al., 2024). Their open-source nature further aligns with scientific principles of transparency and reproducibility, reducing privacy concerns associated with proprietary models like GPT. Although BERT-based models offer numerous advantages for scientific summarization, they struggle with processing lengthy documents due

Proceedings of the 1st Workshop on NLP for Science (NLP4Science), pages 36–46

to token limits. To address this issue, two-stage summarization models have been proposed, where the first stage focuses on identifying and extracting salient sentences or passages, and the second stage on generating a concise and coherent summary based on the extracted sentences (Galanis and Androutsopoulos, 2010; Zhang et al., 2019a; Ma et al., 2021; Rezapour et al.).

In this paper, we build upon the previous work and propose a model for summarizing scientific documents that incorporates the extract-then-compress framework. Our model integrates a "graph-augmented extraction module" that first selects order-based salient sentences from the complete text of long documents in the initial step (see §4.1), and then employs an "abstractive compression module" to generate concise and relevant summaries using the extracted drafts (see §4.2). We test our model on two datasets: First, *SUMPUBMED*, a benchmark dataset for abstractive summarization of biomedical scientific articles (Gupta et al., 2021). Second, *BioConSumm*, a new dataset that we created, which was curated for the purpose of this work and is from the domain of biodiversity conservation. One limitation of existing scientific text summarization tasks is their strong reliance on datasets from well-established domains like computer science and biomedicine, neglecting other research areas. This lack of attention has resulted in a shortage of comprehensive exploration and rich datasets in these underrepresented domains. To address this disparity, we introduce the *BioConSumm* dataset, which also serves as a valuable resource for training and evaluating text summarization models across domains. With the *BioConSumm* dataset, we can investigate the unique challenges and characteristics of summarizing scientific texts in low-resourced fields, assess the effectiveness of existing summarization techniques, and develop domain-specific or domain-agnostic models.

2 Related Work

2.1 Scientific Paper Summarization

Scientific paper summarization involves the generation of a concise summary that captures the essential information and findings of a publication while preserving its core meaning (Yasunaga et al., 2019; An et al., 2021). Automatic text summarization methods can be broadly categorized into two types: extractive (Mihalcea and Tarau, 2004) and abstractive ones (Nenkova and McKeown, 2012). Extractive models focus on identifying the most important information, such as sentences or key phrases, from the original text, and constructing a summary based on these selections. Abstractive models aim to grasp the key ideas from the text and generate new, coherent summaries. Unsupervised, graph-based ranking methods (Mihalcea and Tarau, 2004; Erkan and Radev, 2004) are widely used for extractive summarization. These methods assign weights to sentences in a document using scoring systems like eigenvector centrality or cosine similarity. The sentences with the highest scores are then extracted to form the summaries. Caragea et al. (2014) proposed a key phrase extraction framework that uses a citation network. By identifying important nodes and relationships within the graph, their approach extracts key phrases from scientific papers and incorporates the citation context into the summary. Cohan and Goharian (2015) considered both content and citation contexts for summarizing scientific papers, and showed improved performance over purely content-based methods. Similarly, Abu-Jbara et al. (2013) proposed a hybrid method that uses both citation relationships and text content to summarize scientific papers more effectively. Multi-document summarization techniques have been developed, which combine extractive and abstractive techniques to generate concise summaries from multiple related scientific papers. Yasunaga et al. (2017) proposed a graph-based neural network model for summarizing scientific documents by leveraging multi-document inputs, showing that the combination of citation networks and multiple documents can create more informative summaries. Ding et al. (2014); Ge et al. (2021) incorporated domain-specific ontologies and semantic graphs to enrich the content-based summarization process and showed improved coverage and accuracy in summarizing scientific papers.

Recent improvements in deep learning and neural architectures have resulted in significant improvements in extractive models (Liu, 2019; Nallapati et al., 2017). For instance, BertSumExt (Liu and Lapata, 2019) leverages a decoder and encoder architecture as well as a classifier to select the most salient sentences from a document and generate more coherent summaries compared to the previous models. Liu (2019) introduced BERTSUM, which set new performance benchmarks across domains, including scientific text, by incorporating inter-sentence dependencies and sentence-level classification. Nallapati et al. (2016) developed a

sequence-to-sequence model with attention mechanisms that generate more fluent and human-like summaries by capturing long-range dependencies within the text. Further developments, such as PEGASUS (Zhang et al., 2020) and T5 (Raffel et al., 2020), have expanded on this sequence-to-sequence architecture with large-scale pretraining on text generation tasks, enhancing abstractive summarization capabilities.

2.2 Two-stage Summarization

Hybrid, two-stage document summarization systems combine both extractive and abstractive techniques to improve summary quality (Galanis and Androutsopoulos, 2010; Zhang et al., 2019a; Ma et al., 2021; Rezapour et al.). The first stage typically involves the extraction of salient fragments from the original text as raw summaries. In the second stage, these fragments are arranged into summaries. For example, Chen and Bansal (2018) and Bae et al. (2019) followed a hybrid extract-then-rewrite architecture, with policy-based RL to bring the extraction and rewriting modules together. Lebanoff et al. (2019); Xu and Durrett (2019); Mendes et al. (2019) adopted the extract-then-compress paradigm, namely compressive summarization, which first trains an extractor to select salient sentences that are later input to a decoder to generate the summaries. Gehrmann et al. (2018) first selected key sentences through extractive methods and then rewrote them using abstractive techniques, balancing factual accuracy and fluency. Zhong et al. (2020) proposed a matching-based extractive summarization framework, which can be viewed as an extract-then-match framework. This framework employs a sentence extractor to first generate candidate summaries. It then refines these summaries to align more closely with the source document in the learned semantic space.

3 Data

3.1 Data collection

BioConSumm. Our dataset consists of a collection of academic papers in the domain of biodiversity and bio-conservation. In contrast to high-resourced fields such as biomedicine and computer science, where meticulously curated, high-quality datasets for training models are available, obtaining data for low-resourced domains like biodiversity conservation may require a multi-step approach: We first collected paper records from the Web of Science (WoS). These records include meta-information such as author details, paper titles, and digital object identifiers (DOIs), among other relevant data. Given that WoS has already organized its records according to predefined research domains, including biodiversity conservation, we used the WoS query system to specify the category as biodiversity conservation. As of October 2020, there were over 120,000 records pertaining to journal and conference papers within the field of biodiversity conservation. To narrow down the search, we additionally specified the WoS topic as biodiversity, focusing on papers where the title, abstract, or keywords contained the term "biodiversity." Finally, we downloaded more than 20,000 records as *ciw* files, imported them into Endnote X9[1], and leveraged the Endnote API to find and download full papers in the format of PDF based on the WoS records. Note that the Endnote API is limited to downloading papers from databases that permit automatic downloading and are subscribed to by the researchers' organizations. As a result, we downloaded 12,168 full papers in the format of PDF. Given that the texts in some PDF files use unknown encoding, we extracted texts from 11,579 PDF files as our final dataset. This data collection process can be extended to other domains that have limited resources or lack well-curated computational datasets but for which WoS contains records. While we are not allowed to share the full text of the papers, our data extraction pipeline is reproducible, and using the paper titles and our pipeline, researchers can extract the data.

SUMPUBMED. We used *SUMPUBMED* (Gupta et al., 2021), a dataset for abstractive summarization of biomedical scientific articles built from 33,772 scientific articles in Bio Med Central (BMS), as a point of comparison. *SUMPUBMED* processed these articles by ensuring that the text and abstract in each article have the same named entities. As Figure 1 shows, unlike the abstracts in *BioConSumm*, every sentence in each *SUMPUBMED* abstract must contain biomedical named entities such as gene identifiers (i.e., H2A.Z) that also appear in the processed main body of the same article.

Conducting experiments on these two datasets from different domains enables a more thorough evaluation of the proposed model and allows us to assess how the model's performance is influenced

[1] https://endnote.com/

Conservation

The Chinese government initiated one of the world's largest conservation programs involving agricultural ecosystems with the implementation of the 'Grain for Green' (*Tui Geng Huan Lin*) forest policy between 1999 and 2003. This is the first study to systematically quantify multiple dimensions of biodiversity, phytochemical quality and economic benefits associated with (1) the Grain for Green's tea (*Camellia sinensis*; Theaceae) initiative; (2) the state's previous forest policy involving tea populations in protected areas and; (3) the indigenous tea agro-ecosystems replaced or overlooked by this conservation program. There are several novel and unexpected findings. While forest populations contained the greatest ecological diversity, agro-forests and mixed crop plots were associated with the greatest genetic diversity, phytochemical quality and economic benefits. Indigenous management practices should be incorporated into conservation in China in order to create policies that are more aligned towards biodiversity conservation and sustainable livelihoods while allowing local communities to maintain their cultural identity through agrarian practices.

SUMPUBMED

By comparing H2A.Z binding to global gene expression in budding yeast strains engineered so that normally unstable transcripts are abundant, we show that H2A.Z is required for normal levels of antisense transcripts as well as sense ones. High levels of H2A.Z at antisense promoters are associated with decreased antisense transcript levels when H2A.Z is deleted, indicating that H2A.Z has an activating effect on antisense transcripts. Decreases in antisense transcripts affected by H2A.Z are accompanied by increased levels of paired sense transcripts. The effect of H2A.Z on protein coding gene expression is a reflection of its importance for normal levels of both sense and antisense transcripts. We now find that H2A.Z is also significantly enriched in gene coding regions and at the 3' ends of genes in budding yeast, where it co-localizes with histone marks associated with active promoters. The histone variant H2A.Z, which has been reported to have both activating and repressive effects on gene expression, is known to occupy nucleosomes at the 5' ends of protein-coding genes.

Figure 1: Example summaries in BioConSumm and SUMPUBMED datasets

by some data characteristics via quantitative metrics and human evaluation. For *BioConSumm*, we used the body of raw scientific articles as the input data, while the ground truth is the abstracts of these articles. Similarly, for *SUMPUBMED*, we use the body and the processed version of abstracts, as described above.[2]

4 Methodology

This section presents our proposed framework for long document summarization, which follows the extract-then-compress paradigm. Our model incorporates a *graph-augmented extraction module* that extracts salient sentences from the full text of long documents as drafts, and a subsequent *abstractive compression module* that generates concise and coherent summaries based on the extractive drafts.

4.1 Graph-Augmented Extraction Module

Motivated by prior studies that leveraged sentence graphs based on TF-IDF cosine similarities in summarization tasks (Erkan and Radev, 2004; Yasunaga et al., 2017), we extended this methodology by incorporating SciBERT (Beltagy et al., 2019) and Graph Convolution Network (GCN) (Kipf and Welling, 2017) to build our graph-augmented extraction module.

4.1.1 Graph-Based Encoder

Given a source document represented as a sequence of sentences $S = [s_1, s_2, ..., s_n]$, we construct an undirected sentence graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ denotes the set of nodes comprising the sentences in the source document, and $\mathcal{E}$ indicates the significant similarity between connected sentences. We computed sentence similarity by following Erkan and Radev (2004) to first derive the TF-IDF feature for each sentence and then calculate the cosine similarity between any two sentences. We set a predefined threshold of 0.1 for cosine similarity based on the optimal performance observed in (Erkan and Radev, 2004). If the computed cosine similarity between two sentences exceeded this threshold, an edge was added to link the two sentences. We denote the adjacency matrix of the resulting graph $\mathcal{G}$ as $\mathbf{A} \in \mathbb{R}^{n \times n}$, where n is the number of sentences.

We next obtained the initial node feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where d is the dimension of the feature vector, by leveraging sentence embeddings produced from SciBERT (Devlin et al., 2019). We inserted a special tag `[CLS]` at the beginning of each sentence. The final hidden state that corresponds to `[CLS]` was used as the sentence embedding.

To facilitate the identification of salient content in the source document, we applied GCN on top of the sentence graph. This technique propagates information across nodes based on the graph structure and refines the node representations. Specifically, we performed a symmetric normalization of the adjacency matrix $\mathbf{A}$ as follows:

$$\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}, \quad (1)$$

where $\tilde{\mathbf{A}}$ is the adjacency matrix $\mathbf{A}$ with self-loops such as $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n$ and $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$ such as $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. To propagate information across layers, we use the following rule for each layer of GCN:

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), \quad (2)$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{n \times d^{(l)}}$ is the hidden states for each node in the l-th layer, $d^{(l)}$ is the dimension of hidden states in the l-th layer, σ is a nonlinear activation function such as ReLU(.), and $\mathbf{W}^{(l)} \in \mathbf{R}^{d^{(l)} \times d^{(l+1)}}$ is the weight matrix to be learned in the l-th layer. Particularly, we initialize the hidden states in the first layer as the initial node embedding: $\mathbf{H}^{(0)} = \mathbf{X}$, and the final sentence representations are denoted as $\mathbf{H}^{(L)} \in \mathbb{R}^{n \times d^{(L)}}$.

[2]Both datasets are in English.

4.1.2 Sentence Order-Based Extraction

We treat our *graph-augmented extraction module* as an extractive summarization. The common ground-truth labels were derived from target summaries using a greedy algorithm (Nallapati et al., 2017). However, the resulting labels are typically binary, indicating only whether a sentence should be extracted or not. Consequently, the model is trained to extract sentences as unordered sets, which does not preserve the coherence of the target summaries. We believe that such binary labels can hinder the performance of the subsequent *abstractive compression module* as they fail to consider the order of sentences even if they are correctly extracted as a set.

In order to address this issue, we propose a new labeling method for our extraction module, which produces soft labels that reflect ranked source sentences following the sentence orders in a target summary. We denote the set of ground-truth sentences $G = \{g_1, g_2, ..., g_x\}$ indicating the sentences in a source document that should be extracted, and the target summary as a sequence of sentences $T = [t_1, t_2, ..., t_y]$. We looped through each $t_i \in T$ and found its most similar source sentence in G based on ROUGE-2 such as $g_{i^*} = \underset{g_j \in G}{\text{argmax}}\ \text{ROUGE}(t_i, g_j)$. To reflect the sentence order, we recorded the ROUGE-2 score c_i between the current summary C_i after adding g_{i^*} and the target summary T:

$$C_i = C_{i-1} \cup g_{i^*} \tag{3}$$

$$c_i = \text{ROUGE}(C_i, T) \tag{4}$$

Once we finished the loop, we normalized the ordered sequence of c_i into a predefined range $[l, u]$ in decreasing order such that a source sentence corresponding to a lower-indexed target sentence receives a higher score, and after training, it is expected to be extracted earlier. Lastly, we used label smoothing (Szegedy et al., 2016) to mix the normalized score $\tilde{c}_k$ and the independent ROUGE score calculated between each source sentence and the target summary (Narayan et al., 2018) for all source sentences $s_k \in S$ as the final score:

$$r_k = \epsilon \cdot \tilde{c}_k + (1 - \epsilon) \cdot \text{ROUGE}(s_k, T), \tag{5}$$

where $\epsilon \in [0, 1]$ is an adaptation factor and we set it to a big value, such as 0.9, so that the final scores are mainly based on the normalized scores derived from sentence orders.

During the training, we added an MLP upon the final sentence representation to predict the score: $\hat{y}_k = \text{MLP}(\mathbf{H}_{k,}^{(L)})$, where $\mathbf{H}_{k,}^{(L)}$ means the k-th row of $\mathbf{H}^{(L)}$ representing the final sentence representation for s_k. We adopted cross-entropy loss to calculate the loss and set the minimization of the loss as the training objective:

$$\mathcal{L} = -\sum_{s_k \in S} r_k \cdot \log(\hat{y}_k) \tag{6}$$

4.2 Abstractive Compression Module

While our graph-augmented extraction module effectively compresses long documents into concise summaries, the resulting summaries are extractive in nature, lacking sentence coherence, which diminishes readability and could mislead readers. To address this limitation, we introduce an *abstractive compression module* that converts the extracted salient sentences into complete and coherent summaries. For this purpose, we employed a Transformer-based pre-trained model (Vaswani et al., 2017) and fine-tuned it to generate the target summary given the output from the graph-augmented extraction module as its input.

We explored two widely used models for text generation in our investigation:

- **T5** (Raffel et al., 2020), an encoder-decoder model pre-trained on a diverse set of unsupervised and supervised tasks. Each task is converted into a text-to-text format to facilitate training and inference.
- **BART** (Lewis et al., 2020), a transformer-based denoising autoencoder designed for pre-training sequence-to-sequence tasks.

By leveraging the abstractive model, we aim to transform the extractive summaries into final summaries that are both coherent and informative, enhancing the overall quality and readability of the summarization output.

5 Experiments

In this section, we first introduce implementation details, and then show experimental results from both quantitative and qualitative evaluation.

5.1 Implementation Details

All models were implemented using the PyTorch framework (Paszke et al., 2019) and Huggingface

transformers (Wolf et al., 2020). For the graph-augmented extraction module, we initialized SciBERT with *allenai/scibert-scivocab-uncased*, and built a 2-layered GCN. We set the dimensions of the hidden layer and output layer for GCN as 200 and, 100 respectively. The dimension of the hidden layer of MLP was set to 32 and we used ReLU as the activation function in MLP. This module was optimized by Adafactor (Shazeer and Stern, 2018) with the learning rate of $5e-5$. As for the abstractive compression module, we initialized BART with *facebook/bart-base* and T5 with *t5-base*. During training, we optimized the model with AdamW (Loshchilov and Hutter, 2018) and set the learning rate to $5e-5$.

5.2 Quantitative Evaluation

To evaluate the performance of our proposed model, we conducted experiments on *BioConSumm*, the conservation dataset that we have curated, as well as with *SUMPUBMED*. To evaluate the quality of the generated summaries, we utilized the widely-used ROUGE scores (Lin, 2004), which were assessed against the ground truth data comprising abstracts of scientific articles. Specifically, we calculated the ROUGE-1, ROUGE-2, and ROUGE-L metrics, which offer insights into the generated summaries' quality.

For both evaluations, we compared our model, referred to as **Ordered**, against several baselines:

- **T5**: We follow a similar approach to the BART baseline by fine-tuning the T5 model on the dataset. Similarly, we truncated the input documents to comply with the maximum token limit imposed by T5.
- **BART**: We directly fine-tuned the BART model on the dataset. In this process, we truncated the input documents to fit within the maximum token limit imposed by BART.
- **Unordered**: This baseline shares the same structure as our model, but we trained the extraction module based on the ROUGE score between each sentence in an input document and the target summary as in Narayan et al. (2018).

5.2.1 Results on BioConSumm

Table 1 shows that BART consistently outperformed T5 across all evaluation scenarios in terms of ROUGE-1 and ROUGE-L. This observation aligns with the widespread success of BART for summarization (Zhang et al., 2021; Lewis et al., 2020). Also, all two-staged models outperformed abstractive models, namely T5 and BART. This could be due to the fact that crucial information that needs to be included in the summary may not always be found at the beginning of the scientific documents, unlike documents in other domains such as news articles (Grenander et al., 2019; Xing et al., 2021). By contrast, two-staged models can encode entire documents, which enables them to capture salient content regardless of its position in a document.

Among the two-staged models, those incorporating sentence order-based extraction (referred to as "Ordered") outperformed the models that do not explicitly consider sentence order (referred to as "Unordered"). This finding validates our initial hypothesis that considering sentence order in the extractive module matters, which is often disregarded in traditional extractive summarization. Our proposed method, which incorporates sentence order into the training of the extractive module, preserves sentence order and explicitly models the inherent coherence and structure within a document during the first stage of summarization. We believe that this feature contributes to the better performance of our proposed models; enabling it to better capture the essence and flow of the original content.

Model	R1	R2	RL
T5	42.97	12.16	20.13
BART	43.55	12.97	20.25
Unordered + T5	45.76	14.37	21.89
Unordered + BART	45.96	14.40	21.69
Ordered + T5	**46.18**	15.16	22.10
Ordered + BART	46.09	**15.51**	**22.86**

Table 1: Experimental results on BioConSumm.

Model	R1	R2	RL
T5	42.16	12.38	19.78
BART	44.87	13.83	20.30
Unordered + T5	46.56	15.35	21.25
Unordered + BART	**46.72**	15.85	21.57
Ordered + T5	46.43	15.72	21.42
Ordered + BART	46.55	**15.88**	**21.77**

Table 2: Experimental results on SUMPUBMED.

5.2.2 Results on SUMPUBMED

We conducted additional experiments on *SUMPUBMED* (Table 2). Consistent with

our findings on *BioConSumm*, all order-based two-staged models outperformed the single abstractive models or unordered models in terms of ROUGE-2 and ROUGE-L. This reaffirms the effectiveness of order-based two-staged models for long document summarization. The improvement on *BioConSumm* data is notably larger (specifically for ROUGE-L) than for *SUMPUBMED*. We conducted a human evaluation to investigate the characteristics of summaries and the reason for this difference.

5.3 Human Evaluation

We sampled 10 articles from each of the two datasets and asked four students fluent in English to read the full papers with their abstracts removed. The abstract (ground truth) and the model-generated summaries of each paper were rated by two students with respect to four aspects: Content coverage, Coherence, Hallucination, and Overall quality (Howcroft et al., 2020). The evaluators were unaware of whether they were assessing the original abstract or a model-generated summary during the evaluation process. The four evaluation aspects are further explained below:

- **Content Coverage**: This includes five items, which aim to evaluate how well a model-generated summary or an abstract covers the main points of the corresponding full paper: 1) research background, 2) research questions or goals, 3) methods, 4) findings, and 5) conclusion or discussion.
- **Coherence**: Three coherence items aim to evaluate 1) how logical a summary or abstract is organized (e.g., background → research questions → methods → findings → contribution), 2) whether bullet points/numbering is correctly formed in a reasonable order if applicable, and 3) how fluent the summary reads.
- **Hallucination**: This aspect aims to evaluate whether a summary or abstract contains any information not mentioned in the paper.
- **Overall Quality**: We asked the evaluators to rate a summary or abstract for its overall quality.

We used a rating scale ranging from 1 to 4 (bad, fair, good, and excellent) for all metrics except numbering and hallucination, which were assessed by a yes-or-no evaluation, with 1 representing "yes" and 0 representing "no."

	BioConSumm		SumPubMed	
	A	M	A	M
Background	3.45	3.2	2.88	2.13
Question	3.05	3.28	2.56	2.67
Method	2.76	2.56	2.75	2.67
Finding	3.25	3.1	2.8	2.8
Conclusion	2.9	2.9	2.33	2.1
Organization	3.4	3.45	2.3	2
Numbering	1	1	1	1
Fluency	3.55	3.35	2.5	2.6
Hallucination	0	0.1	0	0
Overall Quality	3.15	2.8	2.3	2.1

Table 3: Human evaluation of abstracts (A) and model-generated summaries (M) for *BioConSumm* and *SUMPUBMED* data. Since Ordered + BART model is consistently the best model in terms of ROUGE-2 and ROUGE-L as well as the human evaluation scores, we only show results of these models.

Table 3 shows the average human ratings for abstracts and model-generated summaries. The abstracts from *SUMPUBMED* got lower ratings than those from *BioConSumm*, particularly for coherence; a metric that represents the logical order of abstracts in the training data and the generated summaries. This is because *SUMPUBMED* processed all indexed articles by ensuring that the named entities in both the text and abstract in each article were the same. Sentences without shared named entities between the text and abstract were removed, resulting in lower coherence. The higher coherence of the raw texts and abstracts from *BioConSumm* dataset likely provides superior information for training a more effective order-based sentence extraction model, resulting in a more substantial improvement of the final model performance represented by ROUGE scores.

6 Conclusions and Future Work

This paper presents a novel dataset for summarizing scientific articles from the domain of biodiversity conservation, which distinguishes it from existing datasets in this field. Additionally, we proposed a two-staged summarization model that employs the "extract-then-compress" approach to effectively summarize lengthy scientific documents. To evaluate the effectiveness of our model, we compared its performance using both our newly introduced dataset and a benchmark summarization dataset

from the biomedical domain. The results demonstrate that our model outperforms well-established summarization methods.

7 Limitations

While our model exhibits promising performance, there are still challenges to address, particularly in modeling cross-domain datasets. Furthermore, it is important to recognize that evaluating the quality of summaries is a complex task that goes beyond the scope of a single metric. While ROUGE scores have been widely used and accepted as a standard evaluation measure in summarization research (Fabbri et al., 2021; Rezapour et al., 2022), they have inherent limitations, e.g., their focus on lexical overlap, which may not fully capture the nuances of semantic salience. Future work should focus on incorporating additional evaluation methods (e.g., Bertscore (Zhang et al., 2019b)) that consider semantic relevance and coherence and provide a more comprehensive assessment of the summaries.

Finally, expanding datasets to low-resourced domains beyond biodiversity conservation can advance summarization techniques across scientific disciplines. In addition, exploring cross-domain summarization tasks, despite limited training data, addresses the challenges of varying terminology and writing styles. Overcoming these challenges enables the wider application of summarization techniques, promoting knowledge dissemination and interdisciplinary research.

8 Ethical Statement

Our dataset consists exclusively of English-language texts, which may introduce limitations in terms of linguistic diversity and inclusivity. We are committed to promoting open and collaborative research practices. While we cannot share the full texts of our new dataset, a list of paper titles and detailed instructions for reproducing our data collection process are available for future research endeavors [3]. Furthermore, in this analysis, we deliberately chose not to employ any LLMs, whether closed- or open-weight, out of respect for the proprietary nature of our data, ensuring that our methods are fully aligned with ethical standards regarding data usage and model selection.

[3] https://github.com/khan1792/BioConSumm

9 Acknowledgement

We gratefully acknowledge the support from the John D. and Catherine T. MacArthur Foundation.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 596–606.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2021. Enhancing scientific papers summarization with citation graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12498–12506.

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sanggoo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 10–20.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Cornelia Caragea, Florin Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1435–1446.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the association for information science and technology*, 65(9):1820–1833.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, et al. 2019. A summarization system for scientific documents. *arXiv preprint arXiv:1908.11152*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Dimitrios Galanis and Ion Androutsopoulos. 2010. An extractive supervised two-stage method for sentence compression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 885–893.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. Baco: A background knowledge-and content-based framework for citing sentence generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1466–1478.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6019–6024.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. Sumpubmed: Summarization dataset of pubmed scientific articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 292–303.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189.

David M Levy. 2008. Information overload. *The handbook of information and computer ethics*, pages 497–515.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Tinghuai Ma, Qian Pan, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2021. T-bertsum: Topic-aware text summarization based on bert. *IEEE Transactions on Computational Social Systems*, 9(3):879–890.

Alfonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André FT Martins, and Shay B Cohen. 2019. Jointly extracting and compressing documents with summary state representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3955–3966.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*, pages 43–76.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rezvaneh Rezapour, Sravana Reddy, Ann Clifton, and Rosie Jones. Spotify at trec 2020: Genre-aware abstractive podcast summarization.

Rezvaneh Rezapour, Sravana Reddy, Rosie Jones, and Ian Soboroff. 2022. What makes a good podcast summary? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2039–2046.

Athar Sefid and C Lee Giles. 2022. Scibertsum: Extractive summarization for scientific documents. In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, pages 688–701. Springer.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Matt White, Ibrahim Haddad, Cailean Osborne, Ahmed Abdelmonsef, Sachin Varghese, et al. 2024. The model openness framework: Promoting completeness and openness for reproducibility, transparency and usability in ai. *arXiv preprint arXiv:2403.13784*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 948–954.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3292–3303.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019a. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. *arXiv preprint arXiv:2109.04609*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*.