A Comparative Study of Multi-document Summarization Techniques

Anushiya Thevapalan
University of Moratuwa, Sri Lanka
anushiya.22@cse.mrt.ac.lk

Nisansa de Silva
University of Moratuwa, Sri Lanka
nisansadds@cse.mrt.ac.lk

Abstract

Multi-document summarization (MDS) is an approach to extracting a concise and coherent summary of information from multiple source documents. This study presents a comparative analysis of MDS techniques, showcasing the progress in the field. Various MDS techniques are analyzed, and their strengths and weaknesses are compared, providing readers with insights to guide their own research directions. Additionally, benchmark datasets and standard evaluation techniques are presented. The experimental results highlight the variability in model performance across different datasets. For instance, the transformer-based PRIMERA model does well on the Multi-News dataset with a ROUGE-1 score of 42.0 but performs less effectively on others. In contrast, the PEGASUS model is more consistent across datasets, while the LED model excels on the BigSurvey-MDS and MS² biomedical datasets. The graph-based model HETERDOCSUMGRAPH outperforms the transformer-based model on the Multi-News dataset with a ROUGE-1 score of 46.05. The HGSum model, which combines transformer and graph techniques, performs best on the Multi-News dataset with a ROUGE-1 score of 50.64. These findings provide a clear overview of the current MDS techniques, highlighting their strengths and effectiveness in different areas.

1 Introduction

Document summarization is a fundamental task in natural language processing that aims to capture the essential information from the large text while preserving the overall meaning. Although single-document summarization has reached a level of maturity, Multi-Document Summarization (MDS) continues to pose significant challenges in the field of Natural Language Processing (NLP). The complexity arises from the need to amalgamate information from various sources, often characterized by conflicting, duplicate, or complementary details (Ma et al., 2022; Ferreira et al., 2014).

Document summarization techniques can be broadly categorized into three types: abstractive summarization, extractive summarization, and hybrid summarization (Afsharizadeh et al., 2022).

1.1 Abstractive Summarization

Abstractive summarization entails generating fresh phrases and sentences that are not directly copied from the source documents yet effectively convey their meaning. This approach to summarization necessitates a thorough comprehension of the text and the capability to produce original sentences that accurately encapsulate the key concepts and information present in the source documents (Figure 1) (Allahyari et al., 2017).

1.2 Extractive Summarization

On the contrary, extractive summarization focuses on choosing and merging significant sentences or phrases from the source documents to create the summary (Allahyari et al., 2017). This method involves identifying the crucial information within the source documents and selecting the most relevant sentences or phrases to construct the final summary (Figure 1). The specific techniques employed for extractive summarization may vary based on the length and complexity of the source documents. For instance, lengthier sources may require advanced techniques to identify the most vital sentences, whereas shorter sources can be summarized more straightforwardly by extracting keywords or phrases (Allahyari et al., 2017).

1.3 Hybrid Summarization

Hybrid summarization is an amalgamation of abstractive and extractive summarization methods (Ma and Zong, 2022; Afsharizadeh et al., 2022). As shown in Figure 1, this approach utilizes abstractive and extractive techniques to produce the final

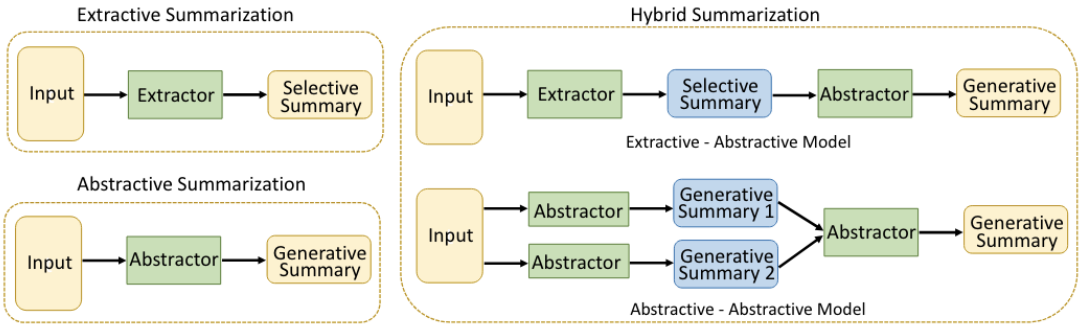

Figure 1: Summarization Construction Types for Text Summarization (Ma et al., 2022)

summary. To illustrate, the system may employ extractive techniques to identify the most crucial sentences or phrases from the source documents and subsequently utilize abstractive techniques to alter or rephrase these sentences for the final summary formation.

Given the diverse approaches to summarization, it is essential to evaluate and compare their effectiveness, particularly in the context of MDS, where challenges can be amplified. We are conducting a comparative analysis of multi-document summarization models to evaluate the performance of transformer-based and graph-based approaches. While transformer-based models have gained significant attention and shown strong results in various summarization tasks, there has been limited exploration of graph-based models in this context. Notably, no previous research has performed a comparative analysis that includes graph-based models alongside their transformer-based counterparts. By including these graph-based techniques in our evaluation, we aim to fill this gap in the literature, providing insights into how different modeling approaches perform across diverse datasets and applications. This analysis will not only enhance our understanding of the strengths and weaknesses of each type of model but also contribute to the development of more effective summarization solutions.

2 Related work

This section presents an overview of the models and techniques employed in this study. It highlights the work carried out by others in various areas relevant to this study.

2.1 Transformer based Methods

PRIMERA (Xiao et al., 2021) is a pre-trained model designed explicitly for Multi-Document Summarization (MDS), extending the capabilities of the Longformer Encoder-Decoder (LED) architecture. The model is pre-trained using a dataset called NewSHead (Gu et al.), which consists of 369,940 clusters of news articles covering similar topics. One of the key innovations in PRIMERA is the use of an Entity Pyramid strategy for generating synthetic summaries during pre-training. This approach identifies essential sentences by analyzing the frequency and relevance of named entities across documents. Through this method, PRIMERA can enhance the quality of summaries by prioritizing the most informative content from multiple sources. Experimental results demonstrate that PRIMERA significantly outperforms earlier MDS models across various benchmarks, showcasing its effectiveness in handling diverse and large-scale multi-document inputs (Xiao et al., 2021).

PEGASUS (Zhang et al., 2020) is a sequence-to-sequence model with gap-sentence generation as a pretraining objective tailored for abstractive text summarization. The key innovation in PEGASUS is its unique pre-training objective, where the model learns to predict entire sentences that have been masked from a document rather than individual tokens, as is common in traditional pre-training approaches like BERT (Zhang et al., 2020). This "gap-sentence generation" (GSG) technique allows PEGASUS to understand better a document's overall structure and key content, which is critical for generating coherent and informative summaries. PEGASUS has been shown to outperform other state-of-the-art models on multiple abstractive summarization benchmarks, particularly excelling in tasks that require summarizing long-form texts (Zhang et al., 2020)

DAMEN (Moro et al., 2022) is a model specifically designed to handle the challenges of multi-

document summarization in the medical field. Unlike standard transformer models that may truncate or overlook critical information, DAMEN addresses this by using a discriminative approach to highlight essential content from clusters of related documents. It leverages token probability marginalization to ensure the generated summaries capture the most relevant details. This method proves particularly effective in the medical domain, where each piece of information is potentially vital (Moro et al., 2022). This approach showcases the potential for tailored neural models in domains requiring high precision, such as healthcare, ensuring the produced summaries are both accurate and informative.

SKT5SciSumm (To et al., 2024) is an Extractive-Generative approach for multi-document scientific summarization. The approach addresses the challenge of processing long and complex scientific texts by combining extractive and abstractive summarization techniques. SKT5SciSumm first utilizes the SPECTER model, which provides sentence-level embeddings trained on scientific texts, to generate dense representations of the input sentences. They then apply K-means clustering to extract the most important sentences from a collection of documents (To et al., 2024). These extracted sentences are passed to the T5 model, a state-of-the-art generative language model, which generates abstractive summaries. This combination balances extracting key information with generating fluent and coherent summaries (To et al., 2024).

2.2 Graph based Methods

In recent years, graph-based approaches have become popular in extractive summarization research. These techniques represent sentences, phrases, or words as nodes in a graph, with the connections between them serving as edges. By analyzing and scoring this network, the models identify and select the most informative and representative sentences to form the summary (Wang et al., 2020; Chen et al., 2021).

Graph-based methods construct graphs of sentences that are part of the document collection. The sentences make the graph's nodes, and edges are either drawn based on the similarity between sentences fulfilling the threshold criteria or belongingness to the same document (Wang et al., 2020; Jiang et al., 2022; Lu et al., 2022). Voting of neighboring nodes selects sentences to generate a summary. In the initial stages of graph-based methods for EDS tasks, researchers primarily utilized similarity scores between sentences in unsupervised ways, employing techniques such as TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004b). The LexPageRank algorithm is based on eigenvector centrality to determine significant sentences, as was done successfully in the Google PageRank algorithm (Erkan and Radev, 2004a).

Data mining techniques have been employed to explore multi-document text summarization. (Baralis et al., 2013) utilized Association Rule Mining within the context of data mining to assess the outcomes of their summarization process. They introduced the GRAPHSUM algorithm to identify correlations among multiple terms in graph-based summarization. The Apriori algorithm was employed for association rule mining to identify correlations among terms, followed by the use of PageRank (Brin and Page, 2012) to rank salient sentences.

Canhasi (2017) introduced a method centered on a Five-Layered Heterogeneous Graph and Universal Paraphrastic Embeddings for query-focused extractive multi-document summarization. This research emphasizes relationships at both the sentence and document levels, incorporating aspects such as part-of-sentence similarity and query-to-sentence similarity (Canhasi, 2017).

Hierarchical graph structures also gained interest in recent research. HAHSum (Jia et al., 2020) is a hierarchical graph designed to address semantic sparsity by leveraging named entities. The HAHSum model incorporates three types of nodes: named entity nodes, word nodes, and sentence nodes. The named entity nodes are represented as anonymized tokens. The graph construction process is as follows: word nodes are connected to the corresponding sentence node with directed edges if they appear within the same sentence. Two named entity nodes are connected by undirected edges if they refer to the same entity, while two sentence nodes are connected by undirected edges if they share a trigram. Furthermore, sequentially occurring words and entities are connected with directed edges. This approach effectively captures implicit information in an explicit manner, allowing for a more comprehensive encoding of the input data (Jia et al., 2020).

HETERDOCSUMGRAPH (Wang et al., 2020) represents cutting-edge developments in graph-based neural networks for summarization tasks. It utilizes heterogeneous graph neural networks to en-

hance extractive summarization. This model constructs a graph where nodes represent sentences and edges capture various relationships such as semantic similarity or document structure (Wang et al., 2020). By leveraging these graph structures, HETERDOCSUMGRAPH can identify and extract the most relevant sentences from multiple documents, ensuring that the resulting summary represents the source material. This graph-based approach allows for a more nuanced understanding of the interconnections between different pieces of information, which is crucial for effective multi-document summarization (Wang et al., 2020).

The development of graph-based summarization techniques marks significant progress in how text is analyzed and summarized. By using graph structures to represent relationships between sentences, words, and entities, these methods provide a strong foundation for creating summaries that are both informative and accurate. As research explores new ways of building and analyzing these graphs, these techniques are expected to become even more critical in creating advanced summarization models that can effectively manage the complexity of today's data.

2.3 Hybrid Methods

CovSumm (Karotia and Susan, 2023) is an unsupervised hybrid summarization model developed for the CORD-19 dataset, which contains COVID-19-related scholarly articles. By combining transformer-based and graph-based methodologies, the model effectively generates summaries from a large volume of scientific texts. The transformer leverages deep learning for contextual understanding, while the graph-based approach enhances the representation of relationships within the documents. This dual strategy enables CovSumm to produce coherent, concise, and informative summaries, making it a valuable resource for researchers and healthcare professionals seeking critical insights from extensive literature.

HGSUM (Li et al., 2023) — an MDS model that extends an encoder-decoder architecture to incorporate a heterogeneous graph to represent different semantic units (e.g., words and sentences) of the documents. To preserve only key information and relationships of the documents in the heterogeneous graph, HGSUM uses graph pooling to compress the input graph (Li et al., 2023).

Khaliq et al. (2024) introduces a framework that integrates topic-aware heterogeneous graph neural networks (HGNN) with transformer models to improve the abstractive summarization of medical scientific documents. This model utilizes HGNN to capture complex relationships among various topics while employing transformer architecture for effective sequence modeling.

3 Datasets

The process of choosing datasets for this study included evaluating their popularity and novelty. We selected datasets from various domains, taking into account the type of data, whether it was short, long, or a hybrid of both document types. Table 4 in the appendix shows the summary of datasets used for evaluation.

The Multi-News dataset (Fabbri et al., 2019) constitutes a comprehensive compilation of news articles and their corresponding human-generated summaries sourced from the website newser.com (Abid, 2022). These summaries are crafted by professional editors and incorporate links to the original articles. One of the most defining characteristics of the Multi-News dataset is its breadth. It incorporates articles from over 1,500 distinct news sites, making it one of the most diverse datasets in the domain of news summarization. It covers various topics, including politics, economics, science, entertainment, etc.

The Multi-News dataset better captures different news organizations' varied styles, tones, and editorial slants, providing a more challenging and realistic task for summarization models. Approximately 85% of the summaries were written by a group of 20 editors, ensuring that the dataset maintains a high level of coherence and consistency across summaries. This contrasts with datasets such as WikiSum (Liu et al., 2018), where the summaries are crowd-sourced or auto-generated from scraped web content, resulting in less uniformity in quality and style. The involvement of professional editors also introduces a diversity of editorial perspectives, as each editor may prioritize different elements of the articles they summarize. This ensures that the dataset encompasses a variety of summarization approaches, which can lead to more robust evaluations of multi-document summarization models. (Abid, 2022).

The Multi-XScience dataset (Lu et al., 2020) is a distinctive compilation formed by merging papers from arXiv.org and Microsoft Academic Graph (MAG) (Sinha et al., 2015) to establish pairs of

target summaries and multi-reference documents. This dataset plays a vital role in addressing the complexities of summarizing scientific literature by pairing target summaries with multiple reference documents from these repositories, allowing for generating summaries that synthesize information across various related papers (Lu et al., 2020).

A vital characteristic of the Multi-XScience dataset is its scale. The dataset consists of 1.3 million arXiv papers. The dataset underwent a curation process involving the cleaning of 1.3 million arXiv LATEX files and aligning them with MAG references. This process was refined through five iterations of cleaning, followed by human validation to ensure data quality (Lu et al., 2020).

The dataset covers various scientific disciplines, including physics, computer science, and biology, making it highly interdisciplinary. This broad scope allows researchers to evaluate how well summarization models generalize across different fields (Lu et al., 2020).

Liu et al. (2018) introduced the WikiSum dataset; generated from summarizing the long Wikipedia articles. It provides a large-scale, naturally occurring dataset. It is built by scrapping the web and designed to address the challenges in long input sequences in multi-document summarization. The dataset includes 1.5 million documents paired with their summaries (Liu et al., 2018).

The BigSurvey dataset (Liu et al., 2023) is introduced in the context of generating structured summaries from large collections of academic papers. This dataset is designed to address the specific challenges associated with summarizing extensive scientific literature. These summaries capture key contributions, methodologies, and results across multiple papers, offering a more organized and systematic representation of the underlying academic content (Liu et al., 2023).

A notable feature of BigSurvey is its emphasis on hierarchical structure, making it well-suited for tasks where organized academic discourse is essential (Li et al., 2023). The summaries are designed to reflect the structure of academic reviews, clearly distinguishing between sections like methods and results—something not addressed in datasets like Multi-News (Fabbri et al., 2019; Li et al., 2023).

Rotten Tomatoes, a website categorizing film reviews, has been widely used in sentiment analysis research. The Rotten Tomatoes dataset (Leone, 2020) includes movie reviews and meta-reviews, with meta-reviews created by professional editors and accompanied by a Tomatometer score, which reflects overall critic reception. This dataset is a key resource for studying sentiment analysis and summarization in the movie review domain.

The Wikipedia Current Events Portal (WCEP) Multi-Document Summarization dataset is a large-scale resource designed to support research in multi-document summarization (MDS) by leveraging a diverse range of real-world events from Wikipedia (Ghalandari et al., 2020). This dataset consists of news articles curated from the Wikipedia Current Events Portal, focusing on global news stories from various domains, such as politics, science, and social events. Each cluster of documents covers a specific event or topic, providing multiple perspectives and details that need to be summarized into a coherent and concise summary (Ghalandari et al., 2020).

4 State-of-the-art Models

We conducted an extensive literature review to identify leading models in multi-document summarization (MDS). From the range of advanced models available, we selected PRIMERA (Xiao et al., 2021), PEGASUS (Zhang et al., 2020), and LED (Beltagy et al., 2020) to represent transformer-based approaches, alongside MGSum (Jin et al., 2020), GraphSum (Li et al., 2020), and HETERDOCSUMGRAPH (Wang et al., 2020) for graph-based models. Additionally, we included HGSum (Li et al., 2023), a hybrid model that combines the strengths of both transformer and graph-based architectures. The selection criteria were based on performance metrics, the year and venue of publication, and the distinct strategies these models employ in processing multi-document inputs. Together, these models encapsulate the current advancements in summarization techniques and provide a broad overview of contemporary methodologies in MDS. Table 3 in the appendix summarises the number of parameters and length of the input and output for the chosen models.

PRIMERA (Pyramid-based Masked Sentence Pre-training for Multi-document Summarization) by Xiao et al. (2021) is noted for its robust performance in handling large-scale document inputs. It leverages a pyramid-based approach to generate summaries by identifying and utilizing hierarchical structures within the text. PRIMERA's ability to manage complex information from multiple sources makes it a strong candidate for this com-

parative analysis.

PEGASUS (Zhang et al., 2020) employs a novel pre-training objective called gap-sentence generation, which has shown state-of-the-art results in various summarization datasets. This model is particularly noted for its strong abstractive summarization capabilities, making it a critical inclusion for evaluating high-quality summarization (Zhang et al., 2020). Similarly, the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) model extends the BERT and RoBERTa architectures to handle longer sequences, addressing the limitations of standard Transformer models in processing long documents effectively.

The Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) model, addresses the challenge of handling long sequences by extending the Transformer architecture. LED uses a combination of local and global attention mechanisms to process long documents without the quadratic increase in computational complexity typical of traditional Transformers (Beltagy et al., 2020). This model is particularly effective in multi-document summarization as it can seamlessly integrate information across extensive texts, ensuring that the summary captures the breadth of content from the input documents (Beltagy et al., 2020).

HETERDOCSUMGRAPH (Wang et al., 2020) utilizes heterogeneous graph neural networks to enhance extractive summarization. This model constructs a graph where nodes represent sentences and edges capture various relationships such as semantic similarity or document structure (Wang et al., 2020). By leveraging these graph structures, HETERDOCSUMGRAPH can identify and extract the most relevant sentences from multiple documents, ensuring that the resulting summary represents the source material. This graph-based approach allows for a more nuanced understanding of the interconnections between different pieces of information, which is crucial for effective multi-document summarization (Wang et al., 2020).

HGSum (Compressed Heterogeneous Graph for Abstractive Multi-Document Summarization) (Li et al., 2023) builds on the graph-based approach by incorporating a hierarchical structure. HGSum introduces a transformer-based approach combining the benefits of graphs. It represents documents at multiple levels of granularity, allowing the model to integrate and synthesize information from various layers (Li et al., 2023). This hierarchical approach ensures that both fine-grained details and high-level summaries are considered, resulting in a comprehensive and coherent summary. HGSum's ability to process and summarize large and complex document sets makes it a strong contender in the field of multi-document summarization (Li et al., 2023).

Including these models in a comparative analysis allows for a thorough examination of the strengths and weaknesses of different approaches to multi-document summarization. This analysis provides a comprehensive overview of the current state-of-the-art in the field by evaluating models that use sparse attention mechanisms, novel pre-training objectives, extended Transformer architectures, and graph-based neural networks. Each model's unique approach to handling the challenges of summarizing multiple documents contributes valuable insights into integrating best and distilling information from diverse sources.

5 Evaluation metrics

ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation, comprises both a set of metrics and a software package employed in the assessment of automatic summarization and machine translation software within natural language processing. These metrics compare automatically generated summaries or translations and a reference (or set of references) provided by humans (Lin, 2004).

BLEU- Bilingual Evaluation Understudy: It employs precision-based metrics to evaluate the similarity between machine-generated and reference translations. The evaluation considers the machine-generated text's fluency (grammatical correctness) and adequacy (semantic equivalence). BLEU is widely utilized in the field of natural language processing and machine translation to quantify the effectiveness of translation models and algorithms. It is used in (Christensen et al., 2013; Tzouridis et al., 2014; ShafieiBavani et al., 2016).

Other evaluation metrics used are Precision, Recall, F-measure, Average Continuity, Pyramid, Correlation coefficients, Amazon mTurk (AMT), etc.

We selected ROUGE for evaluation due to its widespread use in natural language processing, making it easier to compare our results with previous studies. Its broad acceptance ensures consistent benchmarking across various summarization models. By aligning our findings with earlier research, ROUGE enhances the reliability and validity of our comparative analysis.

Datasets	Metric	PRIMERA [1]	PEGASUS [2]	LED [3]	HETERDOCSUMGRAPH [4]
Multi-News [5]	ROUGE-1	42.0 [1]	32.0 [1]	17.3 [1]	46.05 [4]
	ROUGE-2	13.6 [1]	10.1 [1]	3.7 [1]	16.35 [4]
	ROUGE-L	20.8 [1]	16.7 [1]	10.4 [1]	42.08 [4]
Multi-XScience [6]	ROUGE-1	29.1 [1]	27.6 [1]	14.6 [1]	32.56
	ROUGE-2	4.6 [1]	4.6 [1]	1.9 [1]	7.29
	ROUGE-L	15.7 [1]	15.3 [1]	9.9 [1]	17.90
WikiSum [7]	ROUGE-1	28.0 [1]	24.6 [1]	10.5 [1]	30.48
	ROUGE-2	8.0 [1]	5.5 [1]	2.4 [1]	10.04
	ROUGE-L	18.0 [1]	15.0 [1]	8.6 [1]	21.79
BigSurvey-MDS [8]	ROUGE-1	23.9 [11]	38.9 [8]	39.8 [8]	37.26
	ROUGE-2	4.1 [11]	9.0 [8]	9.4 [8]	8.05
	ROUGE-L	11.7 [11]	16.2 [8]	16.1 [8]	15.91
MS^2 [9]	ROUGE-1	12.8 [11]	12.7 [11]	25.8 [11]	23.76
	ROUGE-2	2.0 [11]	1.5 [11]	8.4 [11]	6.92
	ROUGE-L	8.1 [11]	8.3 [11]	19.3 [11]	17.85
Rotten Tomatoes [10]	ROUGE-1	25.4 [12]	27.4 [12]	25.6 [12]	-
	ROUGE-2	8.4 [12]	9.5 [12]	8.0 [12]	-
	ROUGE-L	19.8 [12]	21.1 [12]	19.6 [12]	-
WCEP [13]	ROUGE-1	43.11 [14]	42.43 [14]	43.05 [14]	-
	ROUGE-2	21.85 [14]	17.33 [14]	20.94 [14]	-
	ROUGE-L	35.89 [14]	32.35 [14]	34.99 [14]	-

Table 1: ROUGE Scores of Different Models on Different Datasets. The sources are as follows.
[1] Xiao et al. (2021), [2] Zhang et al. (2020), [3] Beltagy et al. (2020), [4] Wang et al. (2020), [5] Fabbri et al. (2019), [6] Lu et al. (2020), [7] Liu et al. (2018), [8] Liu et al. (2023), [9] DeYoung et al. (2021), [10] Leone (2020), [11] Hewapathirana et al. (2023), [12] DeYoung et al. (2023) [13] Ghalandari et al. (2020) , [14] Li et al. (2023)

6 Experimental results and discussion

This section presents the findings of our study, which compares the performance of state-of-the-art models on different datasets from various domains. Specifically, we report on the performance of these models using ROUGE (Lin, 2004) scores. We also examine the impact of dataset characteristics, such as the number of documents and documents per cluster, on the performance of the models.

We gathered results from previous studies that utilized the same models and datasets. In cases of conflicting findings, we prioritized the most recent study and the conference where it was presented. We also ensured that the original parameters specified in the studies introducing the models were maintained.

In our research, we specifically evaluated the performance of the HETERDOCSUMGRAPH model (Wang et al., 2020) on Multi-XScience (Lu et al., 2020), MS^2 (DeYoung et al., 2021), and on the newly released BigSurveyMDS dataset (Liu et al., 2023). For this evaluation, we used the same parameters as those in the original HETERDOCSUMGRAPH setup and tested the model on the dataset's test set. The results were then compiled and summarized in Table 1 and 2 to facilitate comparison and analysis. To our knowledge, this is the first time the HETERDOCSUMGRAPH model has been evaluated on the Multi-XScience, MS^2, and BigSurvey MDS dataset.

The results from the two tables provide insights into the performance of various models on multiple datasets using the ROUGE evaluation metrics. Across different datasets, the models exhibit varying strengths, highlighting the diversity of approaches in handling multi-document summarization (MDS) tasks.

On the Multi-News dataset, HGSum demonstrated the highest ROUGE-1 score 50.64, significantly outperforming other models like HETERDOCSUMGRAPH 46.05, GraphSum 45.71,

Datasets	Metric	HGSum [1]	GraphSum [2]	MGSum [3]
Multi-News [4]	ROUGE-1	50.64	45.71	45.63
	ROUGE-2	21.69	17.12	16.71
	ROUGE-L	45.90	41.99	40.92
WCEP [5]	ROUGE-1	44.21	39.56	38.88
	ROUGE-2	21.81	14.38	14.22
	ROUGE-L	36.21	29.41	23.37

Table 2: ROUGE Scores of Different Models on Different Datasets. The sources are as follows. All the metrics are taken from [1]

[1] Wang et al. (2020), [2] Li et al. (2020), [3] Jin et al. (2020), [4] Fabbri et al. (2019), [5] Ghalandari et al. (2020) , [6] Li et al. (2023)

and MGSum 45.63. The superior performance of HGSum can be attributed to its hybrid approach, which leverages both graph structures and transformers. Similarly, HGSum excelled in ROUGE-2 and ROUGE-L metrics, indicating its capacity to generate both accurate and coherent summaries from complex multi-document inputs. HETERDOCSUMGRAPH, which relies on graph-based methods, exhibited strong performance, particularly in handling hierarchical sentence relationships. However, transformer-only models such as PRIMERA and PEGASUS performed comparably lower on ROUGE metrics for this dataset.

On the WCEP dataset, HGSum again led the results with the highest ROUGE-1 score (44.21), followed by GraphSum (39.56). Interestingly, HETERDOCSUMGRAPH was not evaluated for this dataset, making it challenging to assess how well graph-based models perform across datasets that focus on different types of input documents. Despite this, the consistent success of hybrid models like HGSum suggests that combining graph structures and transformer models yields superior results when handling real-world, complex datasets.

PEGASUS and LED showed competitive performance when analyzing other datasets like Multi-XScience, WikiSum, and BigSurvey-MDS. For instance, PEGASUS achieved solid results on BigSurvey-MDS (ROUGE-1 of 38.9), though LED marginally outperformed it in ROUGE-1 and ROUGE-2 scores, likely due to its ability to process longer sequences more effectively. However, PRIMERA underperformed on BigSurvey-MDS compared to its performance on datasets like Multi-News and WikiSum, suggesting that some models may be domain-specific.

In the biomedical-focused MS² dataset, LED demonstrated superior performance (ROUGE-1 of 25.8), outperforming both PRIMERA and PEGASUS. This can be attributed to LED's local and global attention mechanism, which is particularly effective in summarizing longer, more detailed texts typical in medical literature. HETERDOCSUMGRAPH also showed strong results in other datasets, but its absence from MS² prevents direct comparison.

In conclusion, hybrid models like HGSum and graph-based models like HETERDOCSUMGRAPH generally outperform traditional transformer-based models across most datasets. This suggests that combining the strengths of both transformers and graph-based approaches can better capture complex relationships between multiple documents, resulting in more accurate and coherent summaries. However, the performance of models can vary significantly across datasets, indicating that dataset characteristics heavily influence model effectiveness.

7 Conclusion

Multi-document summarization (MDS) holds promise in transforming how we process large datasets, but it faces challenges like managing diverse documents, reducing redundancy, and maintaining coherence. This analysis underscores the importance of dataset characteristics in selecting summarization models, as performance varies significantly across domains. There is no universal solution for MDS, and future research should focus on creating more adaptive models capable of handling various dataset features, including document length and topic diversity. Additionally, hybrid approaches and advanced techniques such as transfer learning could improve model robustness and generalizability across different domains.

References

Azal Minshed Abid. 2022. Multi-document text summarization using deep belief network. *International Journal of Advances in Scientific Research and Engineering (IJASRE)*, 8(8):56–65.

Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, Ayoub Bagheri, and Grzegorz Chrupała. 2022. A survey on multi-document summarization and domain-oriented approaches. *Journal of Information Systems and Telecommunication (JIST)*, 1(37):68.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Elena Baralis, Luca Cagliero, Naeem Mahoto, and Alessandro Fiori. 2013. Graphsum: Discovering correlations among multiple terms for graph-based summarization. *Information Sciences*, 249:96–109.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sergey Brin and Lawrence Page. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833.

Ercan Canhasi. 2017. Query focused multi-document summarization based on five-layered graph and universal paraphrastic embeddings. In *Artificial Intelligence Trends in Intelligent Systems: Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol 1 6*, pages 220–228. Springer.

Moye Chen, Wei Li, Jiachen Liu, Xinyan Xiao, Hua Wu, and Haifeng Wang. 2021. Sgsum: transforming multi-document summarization into sub-graph selection. *arXiv preprint arXiv:2110.12645*.

Janara Christensen, Stephen Soderland, Oren Etzioni, et al. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1173.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*.

Jay DeYoung, Stephanie C Martinez, Iain J Marshall, and Byron C Wallace. 2023. Do multi-document summarization models synthesize? *arXiv preprint arXiv:2301.13844*.

Gunes Erkan and Dragomir Radev. 2004a. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 365–371.

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

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Rafael Ferreira, Luciano de Souza Cabral, Frederico Freitas, Rafael Dueire Lins, Gabriel de França Silva, Steven J Simske, and Luciano Favaro. 2014. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, 41(13):5780–5787.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*.

Kushan Hewapathirana, Nisansa De Silva, and C.D. Athuraliya. 2023. Multi-document summarization: A comparative evaluation. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 19–24.

Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3622–3631.

Ming Jiang, Yifan Zou, Jian Xu, and Min Zhang. 2022. Gatsum: graph-based topic-aware abstract text summarization. *Information Technology and Control*, 51(2):345–355.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6244–6254.

Akanksha Karotia and Seba Susan. 2023. Covsumm: an unsupervised transformer-cum-graph-based hybrid document summarization model for cord-19. *The Journal of Supercomputing*, 79(14):16328–16350.

Ayesha Khaliq, Atif Khan, Salman Afsar Awan, Salman Jan, Muhammad Umair, and Megat Farez Azril Zuhairi. 2024. Integrating topic-aware heterogeneous graph neural network with transformer model for medical scientific document abstractive summarization. *IEEE Access*.

Stefano Leone. 2020. Rotten tomatoes movies and critic reviews dataset.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2023. Compressed heterogeneous graph for abstractive multi-document summarization. *arXiv preprint arXiv:2303.06565*.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. *arXiv preprint arXiv:2005.10043*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023. Generating a structured summary of numerous academic papers: Dataset and method. *arXiv preprint arXiv:2302.04580*.

Menghua Lu, Lijia Liang, and Gongshen Liu. 2022. Parallel relationship graph to improve multi-document summarization. In *International Conference on Artificial Neural Networks*, pages 630–642. Springer.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

Ye Ma and Lu Zong. 2022. Parallel hierarchical transformer with attention alignment for abstractive multi-document summarization. *arXiv preprint arXiv:2208.07845*.

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.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 180–189.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2016. On improving informativity and grammaticality for multi-sentence compression. *arXiv preprint arXiv:1605.02150*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Huy Quoc To, Hung-Nghiep Tran, Andr'e Greiner-Petter, Felix Beierle, and Akiko Aizawa. 2024. Skt5scisumm-a hybrid generative approach for multi-document scientific summarization. *arXiv preprint arXiv:2402.17311*.

Emmanouil Tzouridis, Jamal A Nasir, and Ulf Brefeld. 2014. Learning to summarise related sentences. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1636–1647.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

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.

A Model parameter sizes

Model	#parameters	Len-in	Len-out
PRIMERA	447M	4,096	512
PEGASUS	568M	1,024	512
LED	459M	16,384	512
GraphSum	463M	4,050	300
HGSUM	501M	4,096	512
GPT-3	175B	2,049	512

Table 3: Model parameter sizes. Len-in and Len-out denote the maximum lengths of the model input and output, respectively (Li et al., 2023; Floridi and Chiriatti, 2020)

B Summary of datasets used for evaluation

Dataset	Total number of documents	Average number of documents per cluster	Domain
Multi-News (Fabbri et al., 2019)	56K [1]	3.5 [1]	News articles
Multi-Xscience (Lu et al., 2020)	40K [1]	2.8 [1]	Related-work section in scientific articles
Wikisum (Liu et al., 2018)	1.5M [1]	40 [1]	Wikipedia articles
BigSurvey-MDS (Liu et al., 2023)	430K	61.4	Human-written survey papers on various domains
MS^2 (DeYoung et al., 2021)	470K	23.5	Reviews of scientific publications in the medical domain
Rotten Tomato Dataset (Leone, 2020)	244K	26.8	Movie reviews
WCEP (Ghalandari et al., 2020)	650K	63	Human written summaries on news events

[1] (Xiao et al., 2021)

Table 4: Summary of datasets used for evaluation