Multi-document summarization using Graph Neural Networks

229305H - T. Anushiya

Supervisor: Dr. Nisansa de Silva

Content

- Introduction
- Research problem
- Literature review
- Proposed methodology

Introduction

- Multi-document summarization is a vital task in information retrieval, condensing large volumes of text into concise, coherent summaries.
- GNNs excel in capturing the complex relationships between textual elements and provide a powerful framework for representing documents and their connections.

Research problem

Enhancing the Graph Neural Network based multi-document summarization models by introducing novel architecture

The current research offers limited exploration in terms of novel architecture. innovations tailored to enhance the performance of extractive summarization models. This research is expected to substantially elevate the effectiveness of extractive summarization techniques, thus contributing to the broader goal of refining automated text summarization system

Research objectives

- Introduce a benchmark algorithm to exceed the baseline accuracies of existing MDS datasets from [5].
- Introduce a novel architecture by combining existing architectures for MDS and GNN.
- Expand the Heterogeneous graph architecture to improve the performance by utilizing Entity nodes.

Literature Survey

Graph based approaches

- Hetersum
- Heterlongsum
- Hetertreesum

(Elaborate with architectural diagrams)

Literature Survey

Graph Neural Networks (GNNs) for Document Summarization

- Graph Neural Networks (GNNs) is a powerful class of neural networks designed for processing and learning from graph-structured data [13].
- Key Characteristics:
  - Ability to capture complex relationships between interconnected elements.
  - Suited for applications involving structured data, such as document networks.
- Why GNNs for Document Summarization? [2]
  - Graph Representation: Documents can be represented as nodes, with edges denoting semantic relationships.
  - Capturing Context: GNNs excel in capturing the contextual dependencies between sentences and paragraphs in a document.
  - Enhanced Understanding: The ability to model document structure enables more nuanced comprehension, leading to improved summarization.

HetTreeSum: Heterogeneous Tree Structure-based Extractive Summarization Model

Heterogenous Tree structure based extractive summarization to overcome the challenges in scientific paper summarization [8]

1. Inter-sentence relations are hard to learn

2. Structural information of the well-structured scientific papers has not been fully exploited.

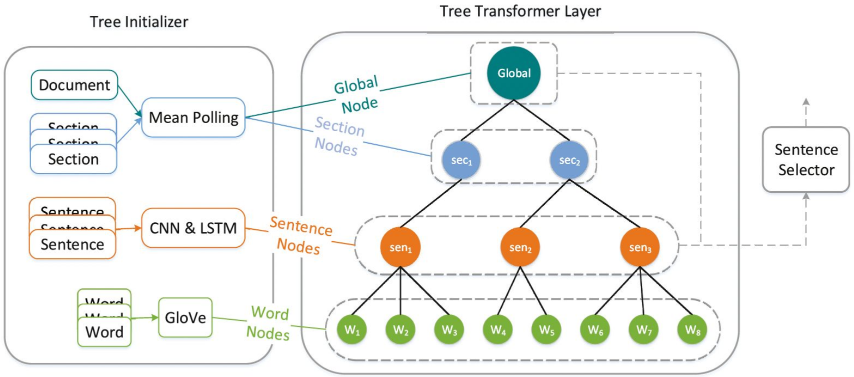


Fig. 1. The whole architecture of HetTreeSum. Units in color green, orange and blue represent for word nodes, sentence nodes and section nodes, respectively. In particular, root of the tree structure in color dark cyan refers to the global node.

Compressed Heterogeneous Graph

- HGSUM is an advanced model extending encoder-decoder architecture with a unique heterogeneous graph approach [6].

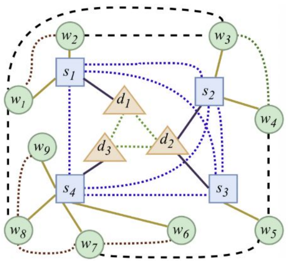
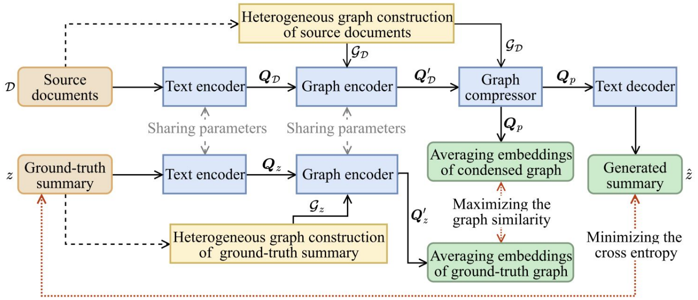


Figure 1: The structure of the heterogeneous graph given three documents in a document cluster: The orange triangles denote document nodes d, the blue quadrates denote sentence nodes s, the green circles denote word nodes w, and the line (or curve) segments between nodes denote edges. A detailed description of the graph is in the Preliminaries.



Heterogeneous graph construction of source documents
$\mathcal{G}_{\mathcal{D}}$
$\mathcal{G}_{\mathcal{D}}$
$\mathcal{D}$
Source documents
Text encoder
$\boldsymbol{Q}_{\mathcal{D}}$
Graph encoder
$\boldsymbol{Q}'_{\mathcal{D}}$
Graph compressor
$\boldsymbol{Q}_p$
Text decoder
Sharing parameters
Sharing parameters
$\boldsymbol{Q}_p$
Averaging embeddings of condensed graph
z
Ground-truth summary
Text encoder
$\boldsymbol{Q}_z$
Graph encoder
Generated summary
$\hat{z}$
$\mathcal{G}_z$
$\boldsymbol{Q}'_z$
Maximizing the graph similarity
Heterogeneous graph construction of ground-truth summary
Averaging embeddings of ground-truth graph
Minimizing the cross entropy

Figure 2: The HGSUM architecture: There are four main components: (1) text encoder (initialised using PRIMERA weights); (2) graph encoder; (3) graph compressor; and (4) text decoder (initialised using PRIMERA weights).

Datasets

- Multinews: Multinews is a dataset designed for multi-document summarization, featuring diverse news articles covering a wide range of topics.
- Multi-XScience: Multi-XScience is a multi-document summarization dataset specifically curated from the field of scientific literature
- WikiSum: WikiSum is a multi-document summarization dataset constructed from Wikipedia articles, offering a mix of topics and genres.

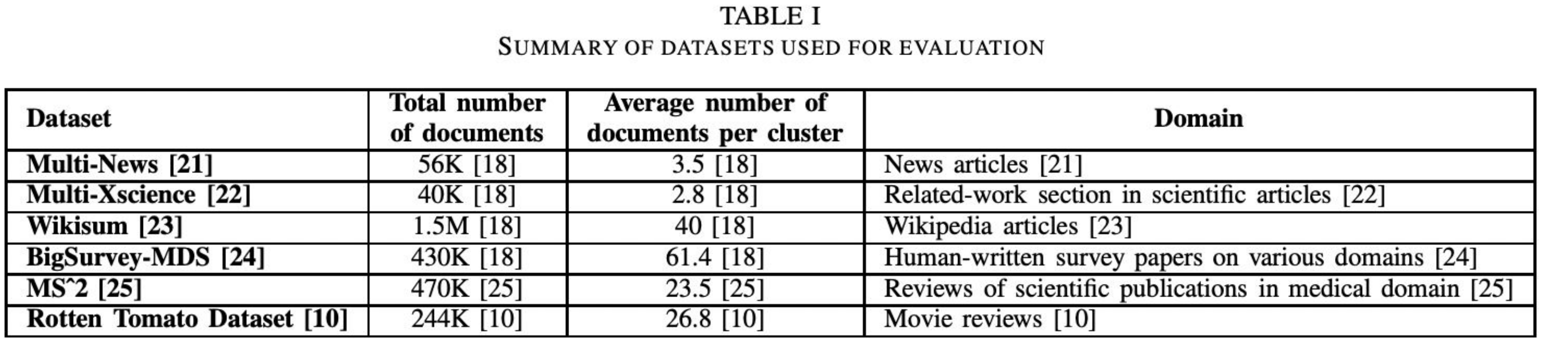


TABLE I
SUMMARY OF DATASETS USED FOR EVALUATION

Dataset	**Total number of documents**	**Average number of documents per cluster**	**Domain**
Multi-News [21]	56K [18]	3.5 [18]	News articles [21]
Multi-Xscience [22]	40K [18]	2.8 [18]	Related-work section in scientific articles [22]
Wikisum [23]	1.5M [18]	40 [18]	Wikipedia articles [23]
BigSurvey-MDS [24]	430K [18]	61.4 [18]	Human-written survey papers on various domains [24]
MS^2 [25]	470K [25]	23.5 [25]	Reviews of scientific publications in medical domain [25]
Rotten Tomato Dataset [10]	244K [10]	26.8 [10]	Movie reviews [10]

Proposed Approach

(Explain the combination approach)

- Draw a diagram of the combined architecture if possible

References

[1] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys, vol. 55, no. 5, pp. 1–37, 2022.

[2] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, B. Long et al., “Graph neural networks for natural language processing: A survey,” Foundations and Trends® in Machine Learning, vol. 16, no. 2, pp. 119– 328, 2023.

[3] X. Liu, Y. Su, and B. Xu, “The application of graph neural network in natural language processing and computer vision,” in 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, 2021, pp. 708–714.

[4] M. Afsharizadeh, H. Ebrahimpour-Komleh, A. Bagheri, G. Chrupala et al., “A survey on multi-document summarization and domain-oriented approaches,” Journal of Information Systems and Telecommunication, vol. 10, no. 37, pp. 68–79, 2022.

[5] K. Hewapathirana, N. De Silva, and C. Athuraliya, “Multi-document summarization: A comparative evaluation,” in 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2023, pp. 19–24.

[6] M. Li, J. Qi, and J. H. Lau, “Compressed heterogeneous graph for abstractive multi-document summarization,”arXiv preprint arXiv:2303.06565, 2023.

References

[7] T.-A. Phan, N.-D. N. Nguyen, and K.-H. N. Bui, “Hetergraphlongsum: Heterogeneous graph neural network with passage aggregation for extractive long document summarization,” in Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 6248–6258.

[8] J. Zhao, L. Yang, and X. Cai, “Hettreesum: a heterogeneous tree structure-based extractive summarization model for scientific papers,” Expert Systems with Applications, vol. 210, p. 118335, 2022.

[9] S. Qi, L. Li, Y. Li, J. Jiang, D. Hu, Y. Li, Y. Zhu, Y. Zhou, M. Litvak, and N. Vanetik, “Sapgraph: Structure-aware extractive summarization for scientific papers with heterogeneous graph,” in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022, pp.575–586.

[10] S. Zesheng and Z. Yucheng, “Topic-selective graph network for topic-focused summarization,” arXiv preprint arXiv:2302.13106, 2023.

[11] T.-A. Phan, N. D. Nguyen, and K.-H. N. Bui, “Extractive text summarization with latent topics using heterogeneous graph neural network,” in Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, 2022, pp. 749–756.

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

References

[14] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, vol. 22, pp. 457-479, 2004.

[15] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004.

[16] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in Proceedings of the 7th International Conference on World Wide Web, 1998.

Multi-document summarization using Graph Neural Networks

229305H - T. Anushiya

Supervisor: Dr. Nisansa de Silva

Content

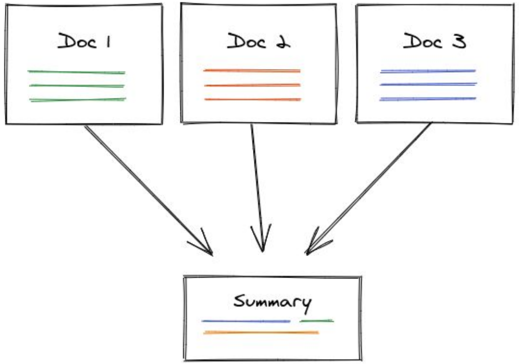
- Introduction
- Research problem
- Literature review
- Datasets
- Proposed methodology

Introduction

- Multi-document summarization is a vital task in information retrieval, condensing large volumes of text into concise, coherent summaries [1]

- Traditional approaches to summarization often struggle to capture the intricate relationships and dependencies present in a network of documents.

- Graph Neural Networks (GNNs) excel in capturing the complex relationships between textual elements and provide a powerful framework for representing documents and their connections. GNNs emerge as a promising solution for document summarization [2].



[1] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys, vol. 55, no. 5, pp. 1–37, 2022.
[2] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, B. Long et al., “Graph neural networks for natural language processing: A survey,” Foundations and Trends® in Machine Learning, vol. 16, no. 2, pp. 119– 328, 2023.

Research problem

Enhancing the Graph Neural Network based multi-document summarization models by introducing novel architecture

Challenges in Traditional Approaches:

- Difficulty capturing complex relationships.
- Limited representation of diverse semantic units.

Graph Neural Networks (GNNs)

- Graph Representation: Documents as interconnected nodes.
- Contextual Understanding: GNNs excel in capturing contextual dependencies.

For multi-document summarization, there are only few research done using Graph Neural Networks (GNNs) so far [7], [21]

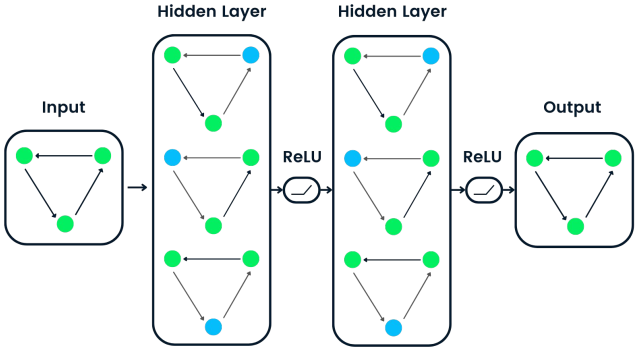
Research objectives

- Introduce a benchmark algorithm to get the baseline accuracies of existing MDS datasets.
- Introduce a novel architecture by combining existing architectures for Multi-document Summarization and Graph Neural Networks.
- Expand the Heterogeneous graph architecture to improve the performance by utilizing Entity nodes.

Literature Survey

Graph Neural Networks

- Graph Neural Networks (GNNs) represent a specialized class of neural networks tailored for processing and analyzing data with inherent graph structures.
- In a graph, entities are represented as nodes, and relationships between them are depicted as edges.
- GNNs excel in capturing complex dependencies and relationships within these graph-structured datasets.



Graph Neural Networks (GNNs) for Document Summarization

- Graph Neural Networks (GNNs) is a powerful class of neural networks designed for processing and learning from graph-structured data [12].
- Key Characteristics:
  - Ability to capture complex relationships between interconnected elements.
  - Suited for applications involving structured data, such as document networks.
- Why GNNs for Document Summarization? [2]
  - Graph Representation: Documents can be represented as nodes, with edges denoting semantic relationships.
  - Capturing Context: GNNs excel in capturing the contextual dependencies between sentences and paragraphs in a document.
  - Enhanced Understanding: The ability to model document structure enables more nuanced comprehension, leading to improved summarization.

[2] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, B. Long et al., “Graph neural networks for natural language processing: A survey,” Foundations and Trends® in Machine Learning, vol. 16, no. 2, pp. 119– 328, 2023.

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

Heterogeneous Graph Neural Network for Extractive Summarisation

- A novel approach to extractive summarization using a heterogeneous graph-based neural network [20].
- This model introduces finer semantic units into the summarization graph, enabling complex relationships between sentences to be captured more effectively.
- This structure can be extended for multi-documents by adding another layer to the graph.

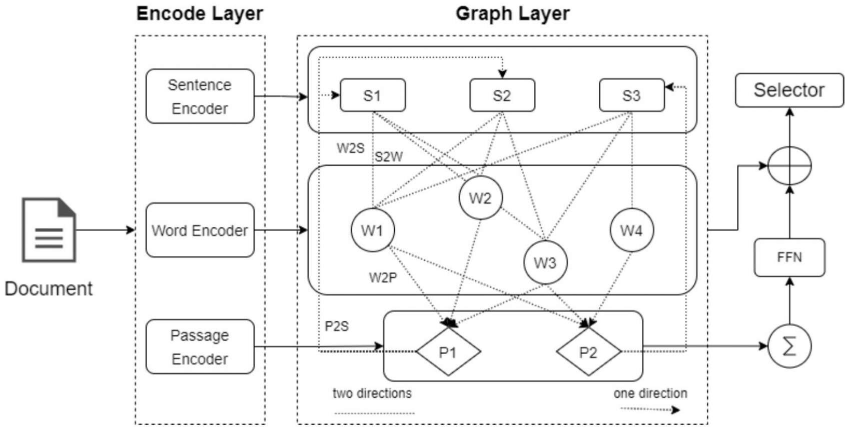


Figure 1: Overview of HeterGraphLongSum model. Passages of each document are defined as a set of sentences in sequence with a fixed number of sentences. In this architecture, the edges from *passage to word* and *sentence to passage* are not taken into account because of the redundancy.

[20]D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, “Heterogeneous Graph Neural Networks for Extractive Document Summarization,” arXiv.org, Apr. 26, 2020.
[7] T.-A. Phan, N.-D. N. Nguyen, and K.-H. N. Bui, “Hetergraphlongsum: Heterogeneous graph neural network with passage aggregation for extractive long document summarization,” in Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 6248–6258.

Heterogeneous Graph Neural Network for Extractive Summarisation

Key components [7]

- **Heterogenous nodes**: The model creates a graph that includes words, sentences, and passages to understand the document better.
- **Node Interaction**: Words help to understand sentences, and sentences help to understand passages, showing how different parts of the text are connected.
- **Passage Summarization**: It focuses on summarizing each passage
- **Selecting Summary Sentences**: Uses information from passages to pick the most important sentences for the summary.

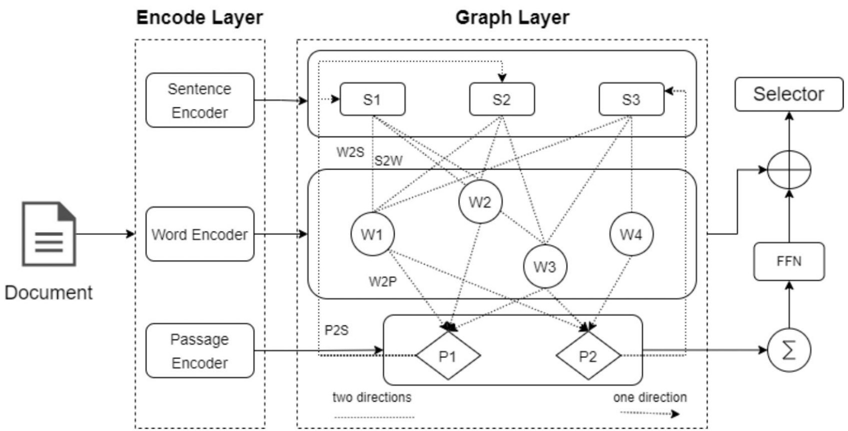


Figure 1: Overview of HeterGraphLongSum model. Passages of each document are defined as a set of sentences in sequence with a fixed number of sentences. In this architecture, the edges from *passage to word* and *sentence to passage* are not taken into account because of the redundancy.

[20]D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, "Heterogeneous Graph Neural Networks for Extractive Document Summarization," arXiv.org, Apr. 26, 2020.
[7] T.-A. Phan, N.-D. N. Nguyen, and K.-H. N. Bui, "Hetergraphlongsum: Heterogeneous graph neural network with passage aggregation for extractive long document summarization," in Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 6248–6258.

Heterogeneous Tree Structure-based Extractive Summarization Model

Heterogenous Tree structure based extractive summarization to overcome the challenges in scientific paper summarization [8]

1. Inter-sentence relations are hard to learn

2. Structural information of the well-structured scientific papers has not been fully exploited.

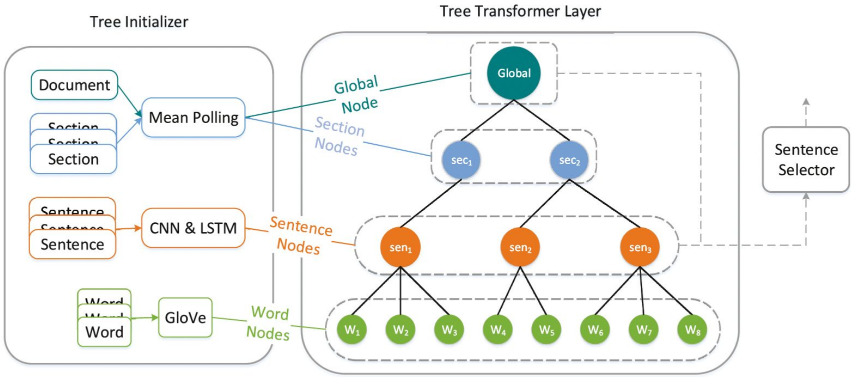


Fig. 1. The whole architecture of HetTreeSum. Units in color green, orange and blue represent for word nodes, sentence nodes and section nodes, respectively. In particular, root of the tree structure in color dark cyan refers to the global node.

[8] J. Zhao, L. Yang, and X. Cai, “Hettreesum: a heterogeneous tree structure-based extractive summarization model for scientific papers,” Expert Systems with Applications, vol. 210, p. 118335, 2022.

Heterogeneous Tree Structure-based Extractive Summarization Model

This architecture comprises three primary components [8]:

1. **Tree Initializer**: Builds a tree structure for the document, setting up nodes for words, sentences, sections, and the entire document.
2. **Tree Transformer Layer**: Iteratively updates each node, using a transformer method to integrate information from connected nodes.
3. **Sentence Selector**: Chooses key sentences for the summary, leveraging the updated node information.

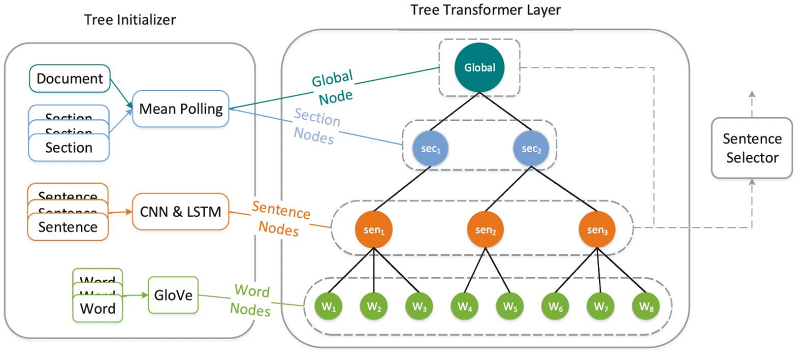


Fig. 1. The whole architecture of HetTreeSum. Units in color green, orange and blue represent for word nodes, sentence nodes and section nodes, respectively. In particular, root of the tree structure in color dark cyan refers to the global node.

[8] J. Zhao, L. Yang, and X. Cai, "Hettreesum: a heterogeneous tree structure-based extractive summarization model for scientific papers," Expert Systems with Applications, vol. 210, p. 118335, 2022.

Compressed Heterogeneous Graph

- A model for multi-document summarization (MDS) that incorporates a heterogeneous graph into an encoder-decoder architecture [6].
- This graph represents various semantic units like words, sentences to enhance the model’s ability to capture diverse relationships in documents [6].
- It's trained with objectives to maximize graph similarity and standard cross-entropy, showing improved performance over existing MDS models

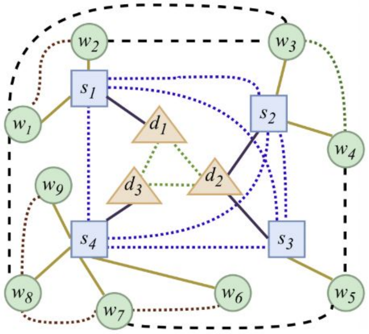


Figure 1: The structure of the heterogeneous graph given three documents in a document cluster: The orange triangles denote document nodes d, the blue quadrates denote sentence nodes s, the green circles denote word nodes w, and the line (or curve) segments between nodes denote edges. A detailed description of the graph is in the Preliminaries.

[6] M. Li, J. Qi, and J. H. Lau, “Compressed heterogeneous graph for abstractive multi-document summarization,”arXiv preprint arXiv:2303.06565, 2023.

Compressed Heterogeneous Graph

Four main component [6]

1. **Text Encoder**: Initializes using PRIMERA weights. It processes the input text to create contextual embeddings.
2. **Graph Encoder**: Utilizes these embeddings to generate graph encodings for the source documents and the ground-truth summary.
3. **Graph Compressor**: Focuses on compressing the graph encoding of the source documents, selecting salient nodes and edges for summarization.
4. **Text Decoder**: Also initialized with PRIMERA weights, it uses the compressed graph encoding to generate the final summary.

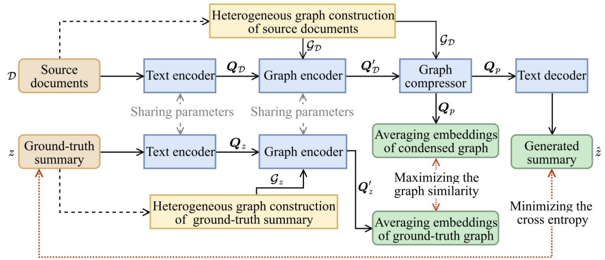

Heterogeneous graph construction of source documents
$\mathcal{G}_\mathcal{D}$
$\mathcal{G}_\mathcal{D}$
$\mathcal{D}$
Source documents
Text encoder
$Q_\mathcal{D}$
Graph encoder
$Q'_\mathcal{D}$
Graph compressor
Q_p
Text decoder
Sharing parameters
Sharing parameters
Q_p
z
Ground-truth summary
Text encoder
Q_z
Graph encoder
Averaging embeddings of condensed graph
Generated summary
$\hat{z}$
$\mathcal{G}_z$
Q'_z
Maximizing the graph similarity
Minimizing the cross entropy
Heterogeneous graph construction of ground-truth summary
Averaging embeddings of ground-truth graph

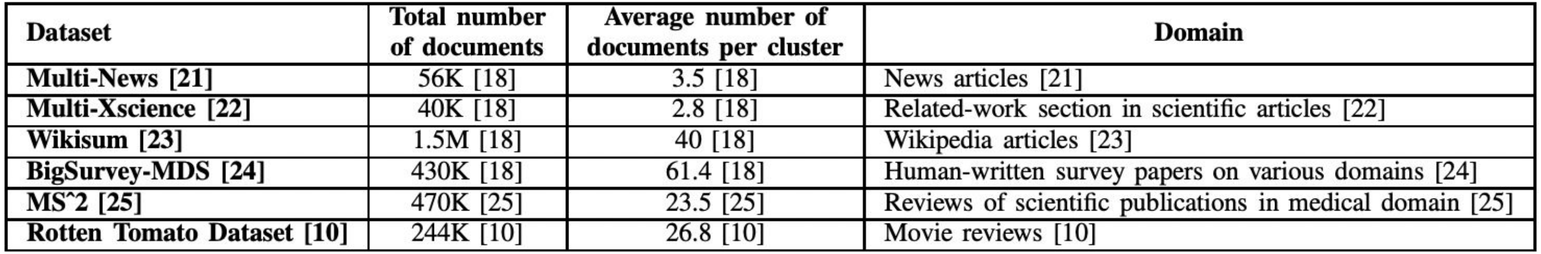
Figure 2: The HGSUM architecture: There are four main components: (1) text encoder (initialised using PRIMERA weights); (2) graph encoder; (3) graph compressor; and (4) text decoder (initialised using PRIMERA weights).

[6] M. Li, J. Qi, and J. H. Lau, “Compressed heterogeneous graph for abstractive multi-document summarization,”arXiv preprint arXiv:2303.06565, 2023.

Datasets

- Multinews: Multinews is a dataset designed for multi-document summarization, featuring diverse news articles covering a wide range of topics [17].
- Multi-XScience: Multi-XScience is a multi-document summarization dataset specifically curated from the field of scientific literature [18].
- WikiSum: WikiSum is a multi-document summarization dataset constructed from Wikipedia articles, offering a mix of topics and genres [19].

TABLE I
SUMMARY OF DATASETS USED FOR EVALUATION [5]

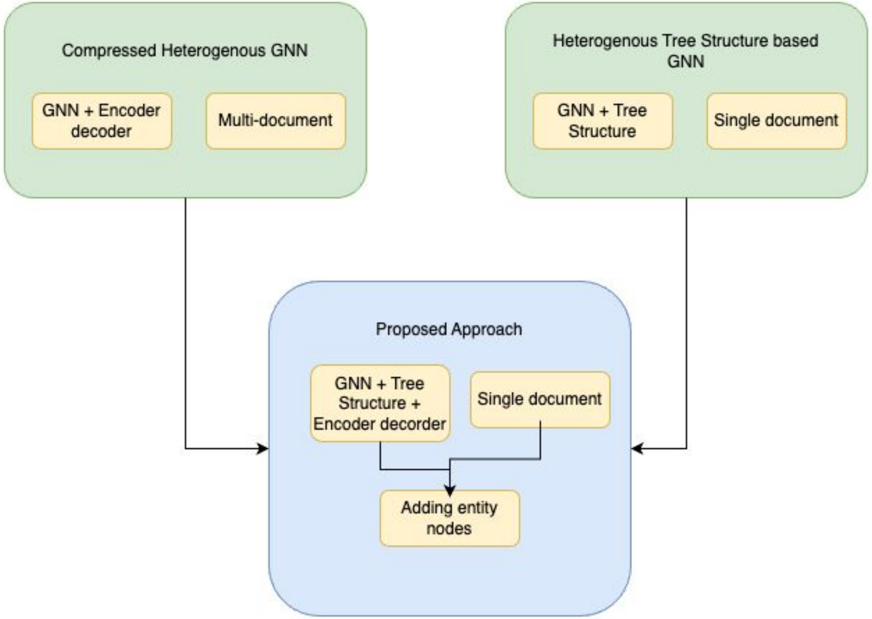


Dataset	**Total number of documents**	**Average number of documents per cluster**	**Domain**
Multi-News [21]	56K [18]	3.5 [18]	News articles [21]
Multi-Xscience [22]	40K [18]	2.8 [18]	Related-work section in scientific articles [22]
Wikisum [23]	1.5M [18]	40 [18]	Wikipedia articles [23]
BigSurvey-MDS [24]	430K [18]	61.4 [18]	Human-written survey papers on various domains [24]
MS^2 [25]	470K [25]	23.5 [25]	Reviews of scientific publications in medical domain [25]
Rotten Tomato Dataset [10]	244K [10]	26.8 [10]	Movie reviews [10]

[5] K. Hewapathirana, N. De Silva, and C. Athuraliya, “Multi-document summarization: A comparative evaluation,” in 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2023, pp. 19–24.

Proposed Approach

- A novel architecture combining the benefits of Tree based Architecture [8] and Compressed encoder-decoder architecture [6].
- Second phase is to add entity nodes to the above architecture.



[6] M. Li, J. Qi, and J. H. Lau, “Compressed heterogeneous graph for abstractive multi-document summarization,”arXiv preprint arXiv:2303.06565, 2023.

[8] J. Zhao, L. Yang, and X. Cai, “Hettreesum: a heterogeneous tree structure-based extractive summarization model for scientific papers,” Expert Systems with Applications, vol. 210, p. 118335, 2022.

References

[1] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys, vol. 55, no. 5, pp. 1–37, 2022.

[2] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, B. Long et al., “Graph neural networks for natural language processing: A survey,” Foundations and Trends® in Machine Learning, vol. 16, no. 2, pp. 119– 328, 2023.

[3] X. Liu, Y. Su, and B. Xu, “The application of graph neural network in natural language processing and computer vision,” in 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, 2021, pp. 708–714.

[4] M. Afsharizadeh, H. Ebrahimpour-Komleh, A. Bagheri, G. Chrupala et al., “A survey on multi-document summarization and domain-oriented approaches,” Journal of Information Systems and Telecommunication, vol. 10, no. 37, pp. 68–79, 2022.

[5] K. Hewapathirana, N. De Silva, and C. Athuraliya, “Multi-document summarization: A comparative evaluation,” in 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2023, pp. 19–24.

[6] M. Li, J. Qi, and J. H. Lau, “Compressed heterogeneous graph for abstractive multi-document summarization,”arXiv preprint arXiv:2303.06565, 2023.

References

[7] T.-A. Phan, N.-D. N. Nguyen, and K.-H. N. Bui, “Hetergraphlongsum: Heterogeneous graph neural network with passage aggregation for extractive long document summarization,” in Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 6248–6258.

[8] J. Zhao, L. Yang, and X. Cai, “Hettreesum: a heterogeneous tree structure-based extractive summarization model for scientific papers,” Expert Systems with Applications, vol. 210, p. 118335, 2022.

[9] S. Qi, L. Li, Y. Li, J. Jiang, D. Hu, Y. Li, Y. Zhu, Y. Zhou, M. Litvak, and N. Vanetik, “Sapgraph: Structure-aware extractive summarization for scientific papers with heterogeneous graph,” in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022, pp.575–586.

[10] S. Zesheng and Z. Yucheng, “Topic-selective graph network for topic-focused summarization,” arXiv preprint arXiv:2302.13106, 2023.

[11] T.-A. Phan, N. D. Nguyen, and K.-H. N. Bui, “Extractive text summarization with latent topics using heterogeneous graph neural network,” in Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, 2022, pp. 749–756.

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

References

[14] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, vol. 22, pp. 457-479, 2004.

[15] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004.

[16] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in Proceedings of the 7th International Conference on World Wide Web, 1998.

[17] A. R. Fabbri, I. Li et al., “Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model,” in ACL, 2019, pp. 1074–1084.

[18] Y. Lu, Y. Dong, and L. Charlin, “Multi-xscience: A largescale dataset for extreme multi-document summarization of scientific articles,” in EMNLP, 2020, pp. 8068–8074.

[19] P. J. Liu, M. Saleh et al., “Generating wikipedia by summarizing long sequences,” in ICLR, 2018.

[20]D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, “Heterogeneous Graph Neural Networks for Extractive Document Summarization,” arXiv.org, Apr. 26, 2020.

Thank You