Multi-Document Summarization: A Comparative Evaluation

Presented by:
Kushan Hewapathirana
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka

Contents

INTRODUCTION

Introduction to Multi-document Summarization

MDS is an automatic process that aims to extract relevant information from multiple texts written about the same topic and represent it in a short piece of text [1]

MDS is a technique that uses sentiment analysis to identify the most important and relevant information from multiple documents and present it in a condensed form [1,2]

MDS aims to summarize large volumes of information quickly and concisely, resulting in increased efficiency, improved decision-making, and better understanding of complex topics [1-3]

Sentiment-based MDS can be used in a variety of applications, such as news aggregation, social media analysis, and market research [1,2]

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[2M. Afsharizadeh, H. Ebrahimpour-Komleh et al., “A survey on multi-document summarization and domain- oriented approaches,” Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] A. M. Abid, “Multi-document text summarization using deep belief network,” 2022.

Techniques and Approaches: MDS

Extractive Summarization

- Contain keywords, phrases, and sentences that are extracted from the source documents. [1,2]

Abstractive Summarization

- Generate precise summaries, including paraphrased sentences and new terms that might not be found in the original documents.[1,2]

Types of Sources

Short - tweets, product reviews, or headlines that convey a smaller amount of information [1,2]

Long - news articles or research papers that contain a large amount of information and detail [1]

Hybrid - scientific summary from a long paper with several short corresponding citations. [1-3]

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[2M. Afsharizadeh, H. Ebrahimpour-Komleh et al., “A survey on multi-document summarization and domain- oriented approaches,” Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] A. M. Abid, “Multi-document text summarization using deep belief network,” 2022.

RESEARCH PROBLEM

Challenges against Multi-document Summarization

- Capturing cross-document and in-document relations
- Avoiding redundancy in the resulting summaries [3]
- Handling multiple languages, cultural contexts [1] and different types of documents (short, long, hybrid)
- Ensuring the summary accurately reflects the tone and sentiment of the original documents [1,3]

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[3] A. M. Abid, “Multi-document text summarization using deep belief network,” 2022.

Research Gap

Current Approaches for Multi-document Summarization

Limited assessment of state-of-the-art multi-document summarization models on diverse datasets across different domains and document types [1-3]

Inability to effectively handle hybrid sources of documents (i.e., a mixture of long and short documents) [1-3]

Lack of understanding about model performance on complex and recently-released datasets, hindering identification of limitations and research directions [1,2]

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] A. M. Abid, "Multi-document text summarization using deep belief network," 2022.

Research Objectives

- To evaluate state-of-the-art MDS models on diverse datasets and identify limitations for future research directions.
- Establish a benchmark for existing models on various datasets.

RELATED WORK

Existing Models for Multi-document Summarization

- Several RNN-based models have been proposed for MDS, including R2N2, STDS, GRU-based encoder-decoder architecture, and RL-MMR [7-10].
- CNNs are effective in various NLP tasks and can be used for semantic and syntactic feature representation in Multi-Document Summarization [11-13].
- Transformer-based models are popular in MDS due to their ability to retain long-range dependencies and parallelization advantage, and they can be divided into three categories[1,11,13]: *Flat Transformer, Hierarchical Transformer, Pre-trained language models*
- Recent studies propose different approaches using Transformer-based models, such as multi-granularity interaction network (MGSum) and Parallel Hierarchical Transformer (PHT) with attention alignment at both the word-level and paragraph-level [11-14].

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[7] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, "Ranking with recursive neural networks and its application to multi-document summarization," in Proceedings of the AAAI conference on artificial intelligence, vol. 29, no. 1, 2015.
[8] X. Zheng, A. Sun, J. Li, and K. Muthuswamy, "Subtopic-driven multi-document summarization," in EMNLP-IJCNLP, 2019, pp. 3153–3162
[9] A. Bražinskas, M. Lapata, and I. Titov, "Unsupervised opinion summarization as copycat-review generation," in ACL, Jul. 2020, pp. 5151–5169
[10] . Mao, Y. Qu, Y. Xie, X. Ren, and J. Han, "Multi-document summarization with maximal marginal relevance-guided reinforcement learning," arXiv preprint arXiv:2010.00117, 2020
[11] . Chen, "Convolutional neural network for sentence classification," Master's thesis, University of Waterloo, 2015
[12] . Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, 2014, pp. 69–78
[13] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang, "Learning summary prior representation for extractive summarization," in ACL, 2015, pp. 829–833
[14] H. Jin, T. Wang, and X. Wan, "Multi-granularity interaction network for extractive and abstractive multi-document summarization," in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 6244–6254

Evaluation Metrics

- ROUGE [1,2,24]
  - ROUGE-N measures n-gram recall while ROUGE-L uses longest common subsequence algorithm which are variants of ROUGE
  - ROUGE-W, ROUGE-S, and ROUGE-SU are extensions of ROUGE-N that incorporate weighting and skip-bigram statistics

Commonly used Datasets in MDS

- DUC and TAC datasets: primarily focused on news articles
- WikiSum [26] dataset: created using Wikipedia articles and their cited sources
- Multi-News [27] dataset: sourced from over 1,500 websites
- WikiHow [28] dataset: extracted from an online knowledge base.
- Rotten Tomatoes dataset: focused on movie reviews and meta-reviews

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2]M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[24] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81
[26]P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," arXiv preprint arXiv:1801.10198, 2018
[27] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-news: A large- scale multi-document summarization dataset and abstractive hierarchical model,"
[28] M. Koupaee and W. Y. Wang, "Wikihow: A large scale text summarization dataset," arXiv preprint arXiv:1810.09305, 2018

METHODOLOGY

Model Comparison

- PRIMERA [1, 2, 22]
  - Prior superior performance in studies, novel Entity Pyramid Gap Sentence Generation (GSG) approach
  - Modified LED architecture; focuses on selecting sentences representing document clusters
  - Entity Pyramid GSG: Masks and generates sentences to form a pseudo-summary
  - Leverages entity frequency, uses Cluster ROUGE for sentence selection
- PEGASUS [18]
  - Outperformed PRIMERA sentiment-wise on Rotten Tomatoes dataset
  - Uses standard Gap Sentence Generation (GSG) for pretraining
- LED [19]
  - Commonly used pre-trained baseline model
  - Longformer-based architecture; efficient local+global attention pattern
  - Linear scalability with input size; LED-base and LED-large sizes available
  - Initialization from BART; adjustable number of hidden layers

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[2]M. Afsharizadeh, H. Ebrahimpour-Komleh et al., “A survey on multi-document summarization and domain- oriented approaches,” Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[18] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in International Conference on Machine Learning. PMLR, 2020, pp. 11 328–11 339
[19] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150, 2020.
[22] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263

Dataset Overview

Datasets	Total number of documents	Average number of documents per cluster	Domain
Multi-News [27]	56K [22]	3.5 [22]	News articles [27]
Multi-Xscience [33]	40K [22]	2.8 [22]	Related-work section in scientific article [33]
Wikisum [26]	1.5M [22]	40 [22]	Wikipedia articles [26]
BigSurvey-MDS [31]	430K [22]	61.4 [22]	Human-written survey papers on various domain [31]
MSˆ2 [32]	470K [32]	23.5 [32]	Reviews of scientific publications in medical domain [32]
Rotten Tomato Dataset [30]	244K [29]	26.8 [29]	Movie reviews [30]

[22] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263
[26]P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, “Generating wikipedia by summarizing long sequences,” arXiv preprint arXiv:1801.10198, 2018
[27] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, “Multi-news: A large- scale multi-document summarization dataset and abstractive hierarchical model,”
[29] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace, “Do multi- document summarization models synthesize?” arXiv preprint arXiv:2301.13844, 2023
[30] S. Leon, “Rotten tomatoes movies and critic reviews dataset,” https://bit.ly/RTdataset, 2020, (Accessed on 06/24/2023)
[31] S. Liu, J. Cao, R. Yang, and Z. Wen, “Generating a structured summary of numerous academic papers: Dataset and method,” arXiv preprint arXiv:2302.04580, 2023.
[32] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, “MSˆ2: Multi-Document Summarization of Medical Studies,” in EMNLP, 2021, pp. 7494–7513.
[33] . Lu, Y. Dong, and L. Charlin, “Multi-xscience: A large- scale dataset for extreme multi-document summarization of scientific articles,” in EMNLP, 2020, pp. 8068–8074

PERFORMANCE EVALUATION

Establish a Benchmark for Existing Models on Various Datasets

Datasets	Metric	Models		
		PRIMERA	PEGASUS	LED
Multi-News	R-1	42.0 [22]	32.0 [22]	17.3 [22]
	R-2	13.6 [22]	10.1 [22]	3.7 [22]
	R-L	20.8 [22]	16.7 [22]	10.4 [22]
Multi-XScience	R-1	29.1 [22]	27.6 [22]	14.6 [22]
	R-2	4.6 [22]	4.6 [22]	1.9 [22]
	R-L	15.7 [22]	15.3 [22]	9.9 [22]
WikiSum	R-1	28.0 [22]	24.6 [22]	10.5 [22]
	R-2	8.0 [22]	5.5 [22]	2.4 [22]
	R-L	18.0 [22]	15.0 [22]	8.6 [22]

[22] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263

Establish a Benchmark for Existing Models on Various Datasets

Datasets	Metric	Models		
		PRIMERA	PEGASUS	LED
BigSurvey-MDS	R-1	23.9	38.9 [31]	39.8 [31]
	R-2	4.1	9.0 [31]	9.4 [31]
	R-L	11.7	16.2 [31]	16.1 [31]
Rotten Tomatoes Dataset	R-1	25.4[29]	27.4 [29]	25.6 [29]
	R-2	8.4 [29]	9.5 [29]	8.0 [29]
	R-L	19.8 [29]	21.1 [29]	19.6 [29]
MS2 Dataset	R-1	12.8	12.7	25.8 [34]
	R-2	2.0	1.5	8.4 [34]
	R-L	8.1	8.3	19.6 [34]

[29] J. DeYoung, S. C. Martinez et al., "Do multi-document summarization models synthesize?" arXiv preprint arXiv:2301.13844, 2023..
[31] S. Liu, J. Cao, R. Yang, and Z. Wen, "Generating a structured summary of numerous academic papers: Dataset and method," arXiv preprint arXiv:2302.04580, 2023
[34] L. L. Wang, J. DeYoung, and B. C. Wallace, "Overview of mslr2022: A shared task on multi-document summarization for literature reviews," in Proceedings of the Third Workshop on Scholarly Document Processing, 2022, pp. 175–180

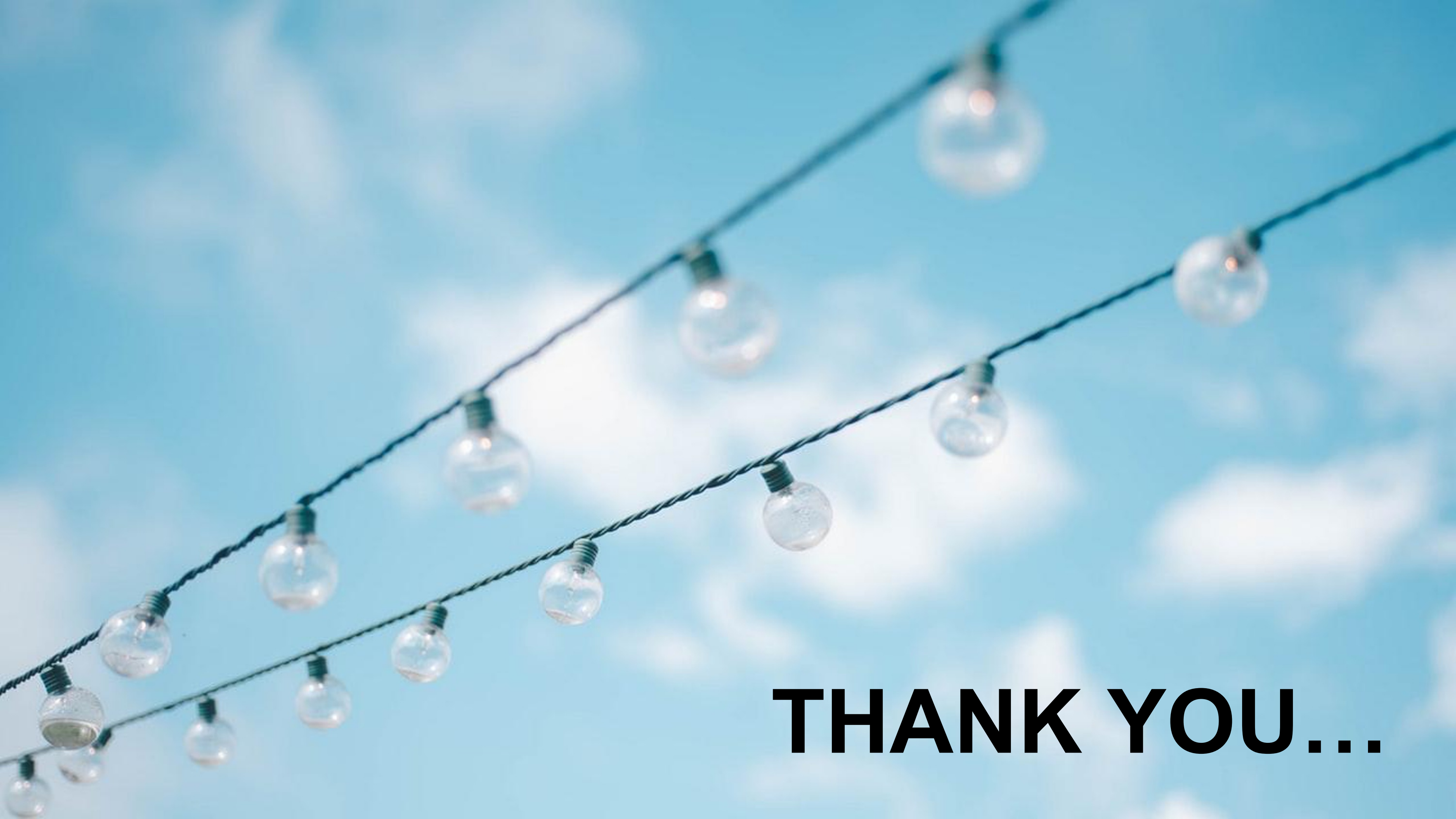
CONCLUSIONS

Conclusions and Future Research Directions

- Conclusions
  - PRIMERA's performance varies across datasets, excelling in Multi-News but struggling in others
  - PEGASUS displays consistent performance across domains, outperforming PRIMERA in some cases
  - LED demonstrates superior performance in the biomedical domain on the MS^2 benchmark
- Insights from Dataset Characteristics
  - Comparisons within the same domain highlight the impact of dataset characteristics
  - PEGASUS and LED perform better than PRIMERA on the BigSurvey-MDS dataset
  - Varying document numbers and documents per cluster affect model performance
- Future Research Directions
  - Overcoming challenges in MDS, including diversity and coherence
  - Enhancing models' domain generalization and adapting to varying dataset characteristics
  - Exploring factors like sentiment for richer summarization dimensions

REFERENCES

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[2M. Afsharizadeh, H. Ebrahimpour-Komleh et al., “A survey on multi-document summarization and domain- oriented approaches,” Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] A. M. Abid, “Multi-document text summarization using deep belief network,” 2022.
[4] C. Ma, W. E. Zhang et al., “Dependency structure for news document summarization,” arXiv e-prints, pp. arXiv–2109, 2021
[5] R. Wolhandler, A. Cattan et al., “How” multi” is multi-document summarization?” arXiv preprint arXiv:2210.12688, 2022.
[6] R. Pasunuru, M. Liu et al., “Efficiently summarizing text and graph encodings of multi-document clusters,” in ACL, 2021, pp. 4768–4779.
7] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, “Ranking with recursive neural networks and its application to multi-document summarization,” in Proceedings of the AAAI conference on artificial intelligence, vol. 29, no. 1, 2015.
[8] X. Zheng, A. Sun, J. Li, and K. Muthuswamy, “Subtopic-driven multi-document summarization,” in EMNLP-IJCNLP, 2019, pp. 3153–3162
[9] A. Bražinskas, M. Lapata, and I. Titov, “Unsupervised opinion summarization as copycat-review generation,” in ACL, Jul. 2020, pp. 5151–5169
[10] . Mao, Y. Qu, Y. Xie, X. Ren, and J. Han, “Multi-document summarization with maximal marginal relevance-guided reinforcement learning,” arXiv preprint arXiv:2010.00117, 2020
[11] . Chen, “Convolutional neural network for sentence classification,” Master’s thesis, University of Waterloo, 2015
[12] . Dos Santos and M. Gatti, “Deep convolutional neural networks for sentiment analysis of short texts,” in Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, 2014, pp. 69–78
[13] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang, “Learning summary prior representation for extractive summarization,” in ACL, 2015, pp. 829–833
[14] H. Jin, T. Wang, and X. Wan, “Multi-granularity interaction network for extractive and abstractive multi-document summarization,” in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 6244–6254
[15] Liu and M. Lapata, “Text summarization with pretrained encoders,” arXiv preprint arXiv:1908.08345, 2019
[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, and others, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” arXiv preprint arXiv:1910.13461, 2019
[17]C. Raffel, N. Shazeer, A. Roberts, K. Lee, and others, “Exploring the limits of transfer learning with a unified text-to-text transformer,” The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020
[18] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in International Conference on Machine Learning. PMLR, 2020, pp. 11 328–11 339
[19] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150, 2020.
[20] M. Zaheer, G. Guruganesh, K. A. Dubey,,., “Big bird: Transformers for longer sequences,” Advances in neural information processing systems, vol. 33, pp. 17 283–17 297, 2020
[21]A. Caciularu, A. Cohan, I. Beltagy, M. E. Peters, A. Cattan, and I. Dagan, “Cdlm: Cross-document language modeling,” arXiv preprint arXiv:2101.00406, 202
[22] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263
[23] G. Moro, L. Ragazzi, L. Valgimigli, and D. Freddi, “Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature,” in ACL 2022, pp. 180–189
[24] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81
[25] Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318
[26]P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, “Generating wikipedia by summarizing long sequences,” arXiv preprint arXiv:1801.10198, 2018
[27] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, “Multi-news: A large- scale multi-document summarization dataset and abstractive hierarchical model,”
[28] M. Koupaee and W. Y. Wang, “Wikihow: A large scale text summarization dataset,” arXiv preprint arXiv:1810.09305, 2018
[29] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace, “Do multi- document summarization models synthesize?” arXiv preprint arXiv:2301.13844, 2023
[30] S. Leon, “Rotten tomatoes movies and critic reviews dataset,” https://bit.ly/RTdataset, 2020, (Accessed on 06/24/2023)
[31] S. Liu, J. Cao, R. Yang, and Z. Wen, “Generating a structured summary of numerous academic papers: Dataset and method,” arXiv preprint arXiv:2302.04580, 2023.
[32] L. L. Wang, J. DeYoung, and B. C. Wallace, “Overview of mslr2022: A shared task on multi-document summarization for literature reviews,” in Proceedings of the Third Workshop on Scholarly Document Processing, 2022, pp. 175–180
[33] . Lu, Y. Dong, and L. Charlin, “Multi-xscience: A large- scale dataset for extreme multi-document summarization of scientific articles,” in EMNLP, 2020, pp. 8068–8074
[34] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, “MSˆ2: Multi-Document Summarization of Medical Studies,” in EMNLP, 2021, pp. 7494–7513.


THANK YOU…