

A decorative background featuring a network diagram with nodes and lines. The nodes are represented by circles of varying sizes and colors (blue, grey, white), connected by thin lines. The diagram is positioned in the top-left and bottom-right corners of the slide.

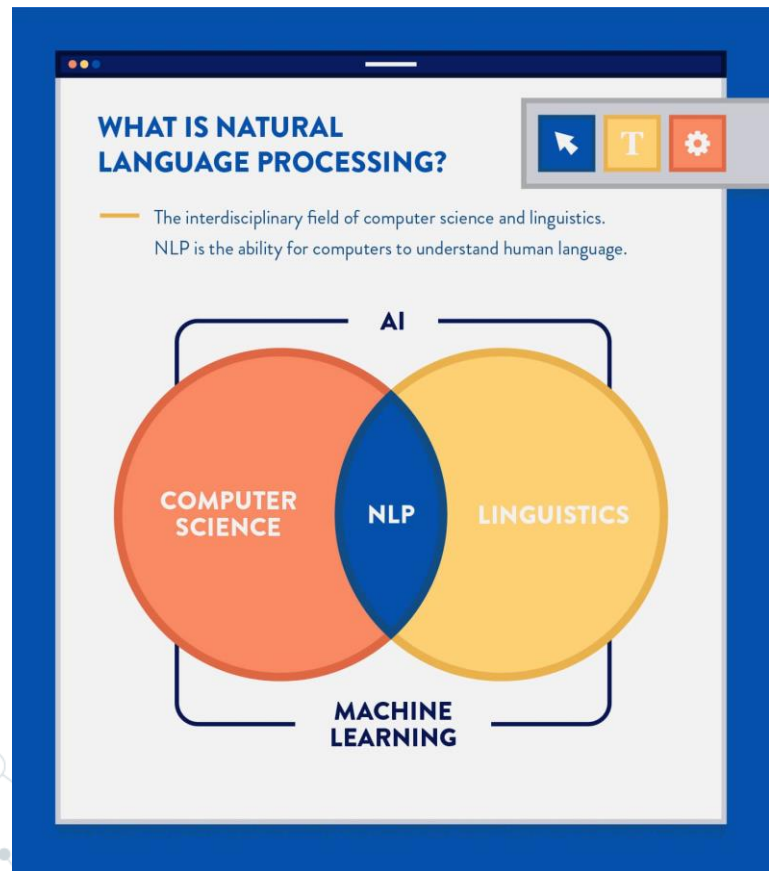
A Novel Approach to Multi-Document Summarization with Sentiment Analysis

Presented by:
Kushan Hewapathirana – 229333P



APPLICATION DOMAIN: Natural Language Processing – Multi-document Summarization

Introduction to Natural Language Processing



**Speech
recognition**

**Part of speech
tagging**

**Word sense
disambiguation**

**Named entity
recognition**

**Co-reference
resolution**

**Sentiment
analysis**

**Natural
language
generation**

Content



Introduction



Research
Problem



Literature
Survey



Progress

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by circles of varying sizes, some with concentric rings, and the lines are thin and grey. The diagram is partially cut off by the left edge of the frame.

INTRODUCTION



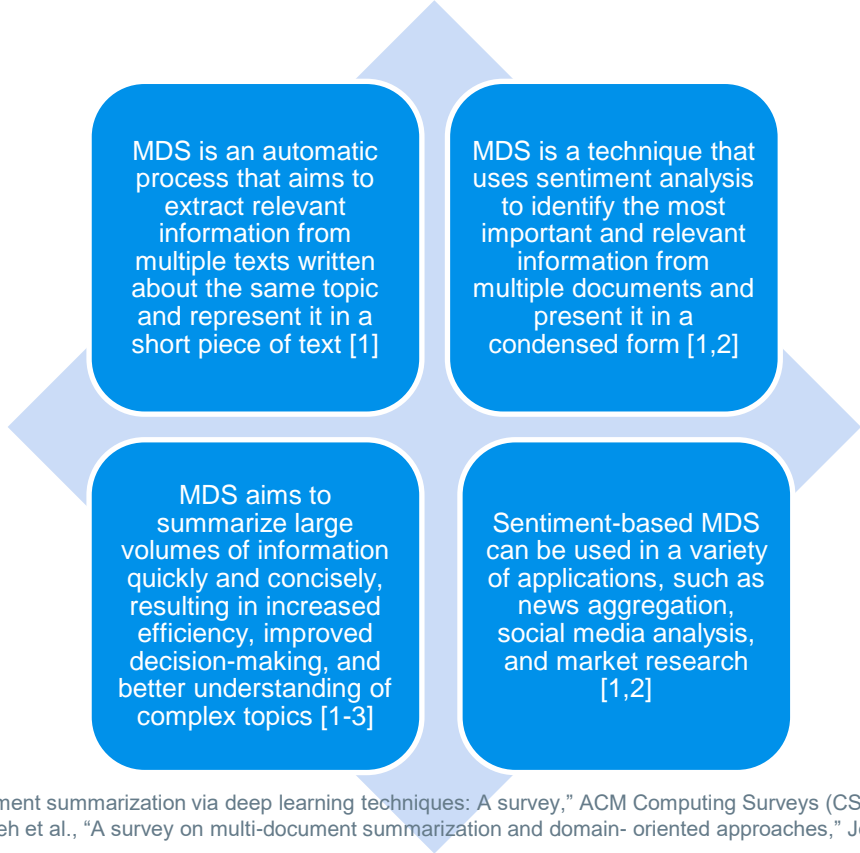
SDS to MDS

- ◎ In terms of the number of texts, the summarization is divided into two categories [1]:
 - Single document summarization (SDS)
 - Multi-document summarization (MDS)
- ◎ Use cases for MDS
 - When a user searches for a topic online, related documents are retrieved, many of which contain similar information
 - An app developer might get multiple reviews of his app
 - A lecturer might get feedback write ups from students
 - A business leader might need to summery of the same news reported at different sources.
- ◎ Single-document summarizers cannot achieve the goal of producing a summary with minimum redundancy and maximum relevance, so multi-document summarization has become a viable solution [1,2].

[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

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

[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

[3] A. M. Abid, "Multi-document text summarization using deep belief network," 2022.

Techniques and Approaches

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

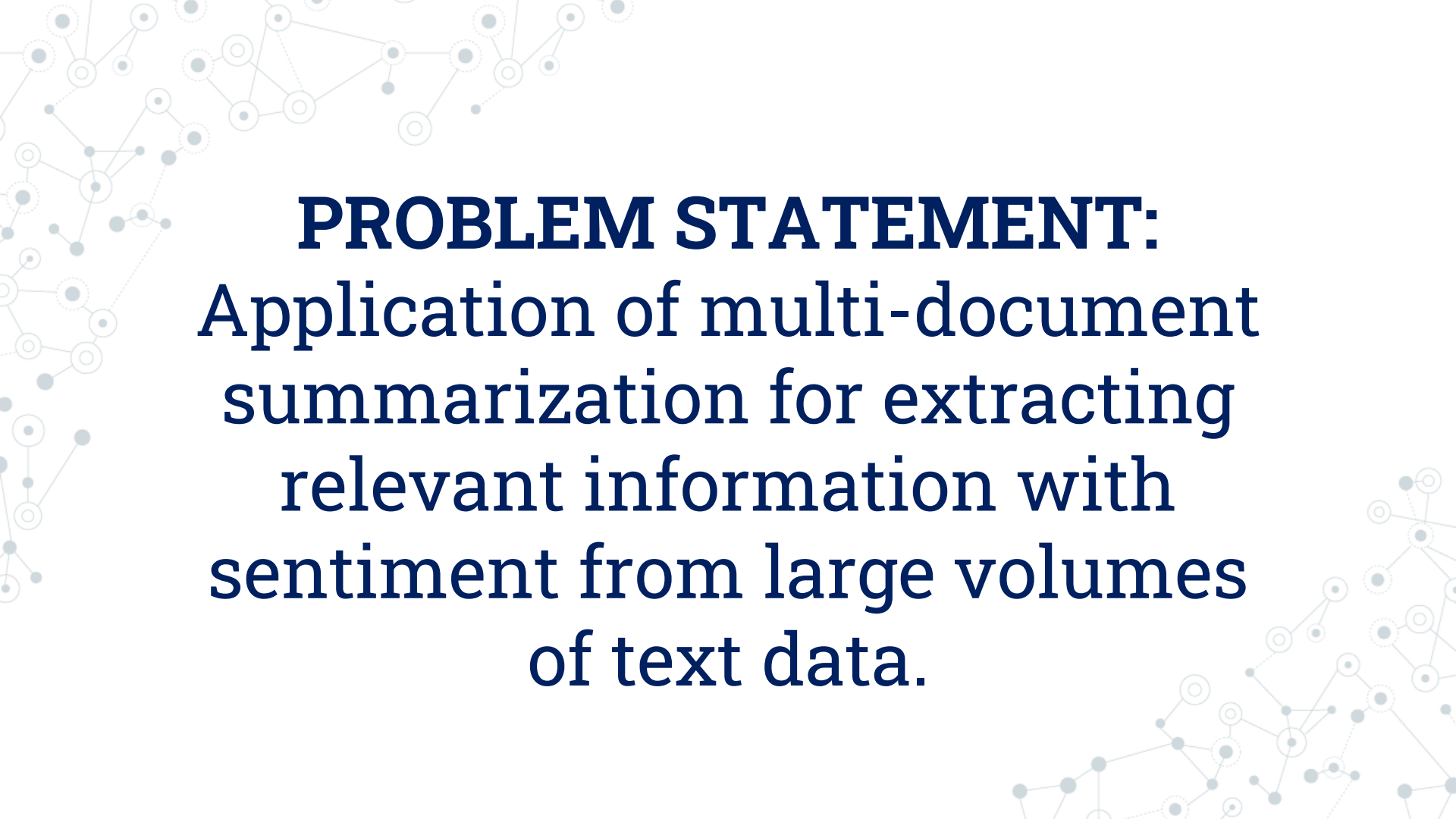
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain-oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68, 2022

[3] A. M. Abid, "Multi-document text summarization using deep belief network," 2022.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles, suggesting a hierarchical or central structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

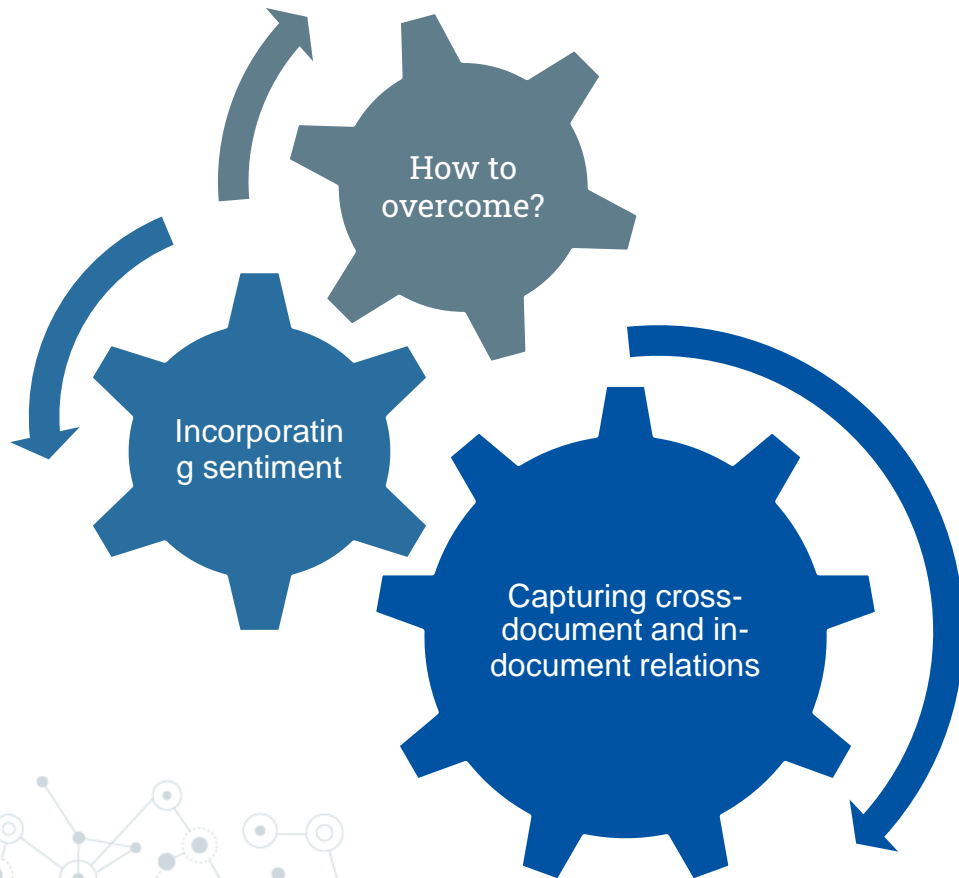
RESEARCH PROBLEM





PROBLEM STATEMENT:
Application of multi-document
summarization for extracting
relevant information with
sentiment from large volumes
of text data.

Challenge against Multi-document summarization



- Avoiding redundancy in the resulting summaries [3]
- Handling multiple languages and cultural contexts [1]
- 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.

Approaches for Multi-document summarization

Limited research on incorporating semantic meaning and context into the summary [1-3]

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

Limited research on adding sentiment annotation to generated summaries [1,2]

[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

[3] A. M. Abid, "Multi-document text summarization using deep belief network," 2022.

Research Objectives

- ① Introduce a benchmark algorithm to establish the baseline accuracies of existing MDS datasets
- ① Create a new MDS dataset which also consists sentiment annotations
- ① Introduce a novel model which incorporates sentiment into the MDS process.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some solid and some hollow, connected by thin lines. The overall structure is organic and branching, resembling a molecular or biological network.

LITERATURE SURVEY



Traditional Summarization Techniques

- ◎ Abstractive summarization involves generating new phrases and sentences that are not present in the source documents, but still capture the meaning of the text [1-4]
- ◎ Extractive summarization, on the other hand, involves selecting and combining important sentences or phrases from the source documents to form the final summary [1,2,5]
- ◎ Hybrid summarization is a combination of abstractive and extractive summarization techniques [1,5]
 - For example, the system may use extractive techniques to identify the most important sentences or phrases from the source documents, and then use abstractive techniques to modify or rephrase these sentences to form the final summary [5]

[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

[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. Cattani et al., "How multi is multi-document summarization?" arXiv preprint arXiv:2210.12688, 2022.

Concatenation Methods

- ◎ One of the key challenges MDS is to merge and concatenate several input documents to produce a unified and coherent summary effectively [1,2]
- ◎ Types of concatenation
 - Flat concatenation [1-3]
 - This method involves simply concatenating all input documents into a single long document, which is then processed and summarized
 - Hierarchical concatenation [1-3, 6]
 - Document-level Concatenation
 - The input documents within a cluster are condensed into extractive or abstractive summaries, or into some form of representation, which are then fused in the subsequent processes for final summary generation
 - Word/Sentence-level Concatenation
 - Sentences within the input documents can be replaced by words, and the relationships among the words/sentences are modeled using clustering algorithms and graph-based techniques

[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

[3] A. M. Abid, "Multi-document text summarization using deep belief network," 2022.

[6] R. Pasupuru, M. Liu et al., "Efficiently summarizing text and graph encodings of multi-document clusters," in ACL, 2021, pp. 4768–4779.

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].

These models aim to address challenges in MDS such as ranking sentence importance, incorporating subtopic information, minimizing diversity of opinions, and including relevance measures.

CNNs are effective in various NLP tasks and can be used for semantic and syntactic feature representation in Multi-Document Summarization [11-13].

Most CNN-based MDS models use multiple filters with different window sizes over the input documents for semantic representation and max-over-time pooling to extract salient feature representation. Some examples of such models are PriorSum, HNet, DPP, MV-CNN, and TCSum [11-14].

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, to address challenges facing abstractive multi-document summarization, including document representation and source coverage [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] J. 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] J. Chen, "Convolutional neural network for sentence classification," Master's thesis, University of Waterloo, 2015

[12] J. 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

Pre-trained Language Models

- ◎ Pre-trained models like BERTSUM [15], BART [16], T5 [17], PEGASUS [18], Longformer [19], and BigBird[20] are effective for summarization tasks.
- ◎ CDLM [21] is a recent work that pre-trains Longformer for cross-document tasks, but only addresses encoder-specific tasks and is not suitable for generation and PRIMERA[22] based on the pre-training of the LongFormer Encoder-Decoder (LED) architecture [19].
- ◎ DAMEN [23] is a method for MDS in the medical domain, which is based on a combination of three language models:
 - Indexer
 - Encodes the background information and documents in a cluster, resulting in dense embedding representations.
 - Discriminator
 - Selects the top k documents through a comparison with the background.
 - Generator
 - Background is combined with the retrieved documents and passed to BART, to generate the multi-document summary.

[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. Cattani, and I. Dagan, "CdLm: Cross-document language modeling," arXiv preprint arXiv:2101.00406, 2021

[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

Evaluation Metrics



ROUGE [1,2,28]

- Measures overlap between generated and reference summaries
- Calculates score based on recall and precision of the generated summary
- 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



BLUE [29]

- Measures n-gram overlap between generated and reference summaries
- Gives higher score for more matching n-grams and penalize for incorrect word order



Other metrics [1,2]

- Precision: measures the proportion of generated summary that is relevant to the reference summary
- Recall: measures the proportion of the reference summary that is covered by the generated summary
- Pyramid: evaluates summaries based on how many content units they cover, where content units are defined based on their position in the source document.

[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

[28] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81

[29] 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

Commonly used datasets in MDS

- ◎ DUC and TAC datasets are primarily focused on news articles and are relatively small and biased towards the first sentence of news articles, while WikiSum, Multi-News, and WikiHow datasets cover a wider range of topics and are larger and more diverse [1,2].
- ◎ WikiSum [24] dataset is created using Wikipedia articles and their cited sources, while Multi-News [25] dataset is sourced from over 1,500 websites, and WikiHow [26] dataset is extracted from an online knowledge base.
- ◎ Multi-News and WikiHow datasets offer an opportunity for models to learn from multiple source documents and summaries, while DUC and TAC datasets focus on single-document summarization[1,2,24-26].
- ◎ The Rotten Tomatoes dataset is focused on movie reviews and meta-reviews, while the other datasets are more general in their scope [27].

[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] 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

[25] 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,"

[26] M. Koupaei and W. Y. Wang, "Wikihow: A large scale text summarization dataset," arXiv preprint arXiv:1810.09305, 2018

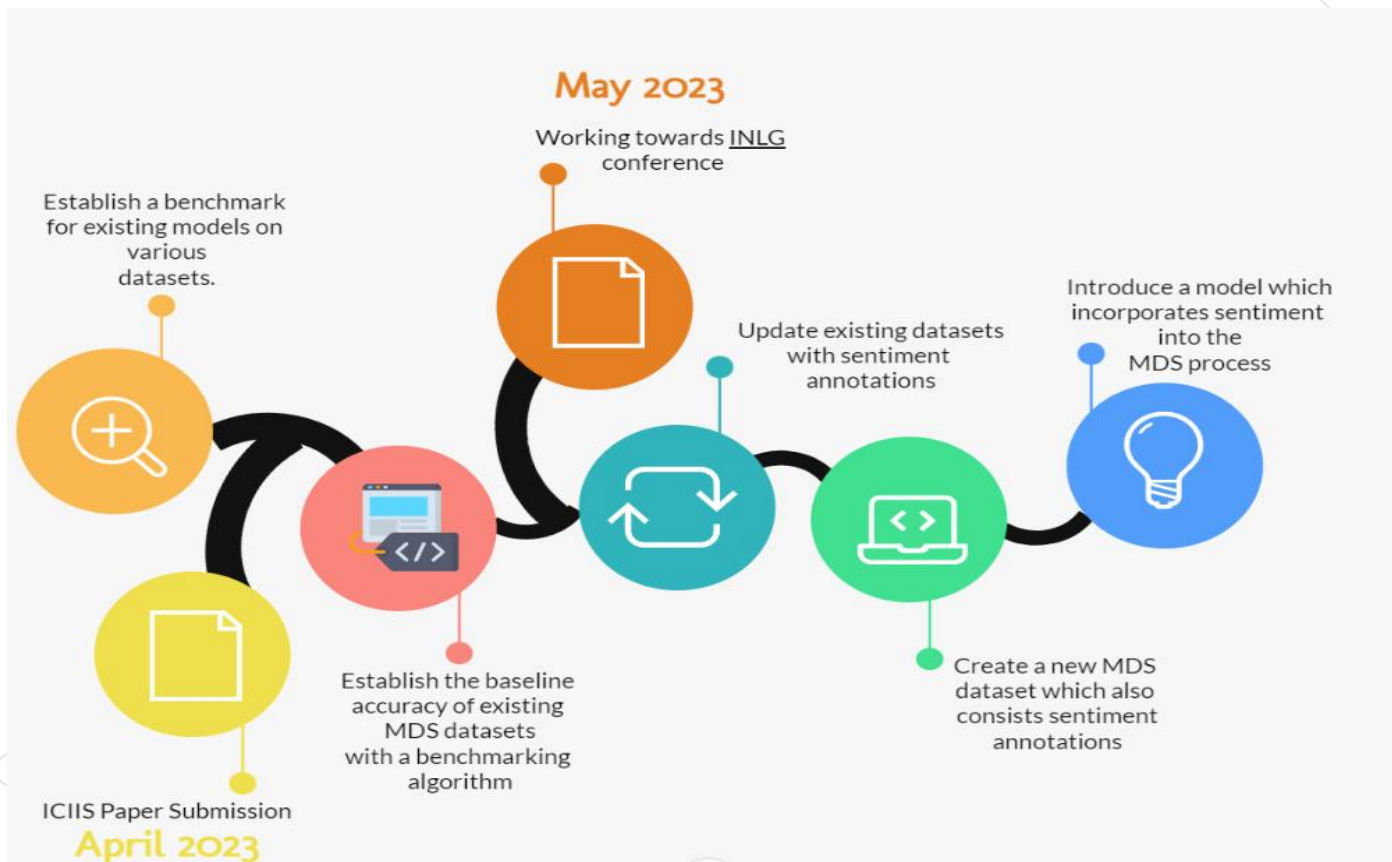
[27] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace, "Do multi-document summarization models synthesize?" arXiv preprint arXiv:2301.13844, 2023

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are solid grey and others are hollow with a grey outline. The lines connecting them are thin and grey, creating a dense, organic structure.

PROGRESS



Planned Methodology



Establish a benchmark for existing models on various datasets

Models	Multi-News			Multi-XScience			WikiSum			BigSurvey-MDS			Rotten Tomatoes Dataset		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
PRIMERA	42.0 [22]	13.6 [22]	20.8 [22]	29.1 [22]	4.6 [22]	15.7 [22]	28.0 [22]	8.0 [22]	18.0 [22]	23.9	4.1	11.7	25.4	8.4 [27]	19.8 [27]
PEGASUS	32.0 [22]	10.1 [22]	16.7 [22]	27.6 [22]	4.6 [22]	15.3 [22]	24.6 [22]	5.5 [22]	15.0 [22]	38.9 [2]	9.0 [30]	16.2 [30]	27.4 [3]	9.5 [27]	21.1 [27]
LED	17.3 [22]	3.7 [22]	10.4 [22]	14.6 [22]	1.9 [22]	9.9 [22]	10.5 [22]	2.4 [22]	8.6 [22]	39.8 [2]	9.4 [30]	16.1 [30]	25.6 [3]	8.0 [27]	19.6 [27]

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

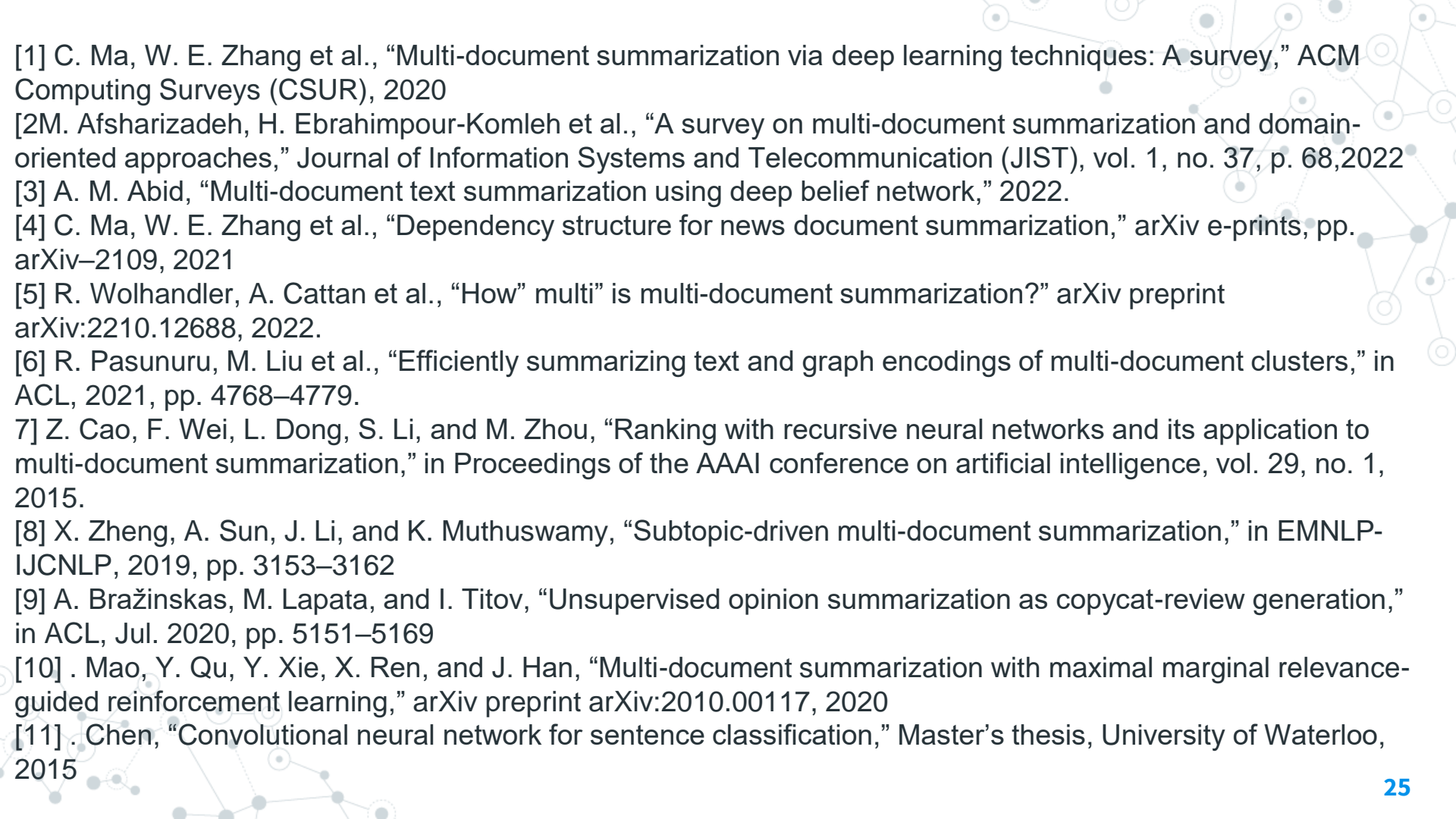
[27] J. DeYoung, S. C. Martinez et al., "Do multi-document summarization models synthesize?" arXiv preprint arXiv:2301.13844, 2023..

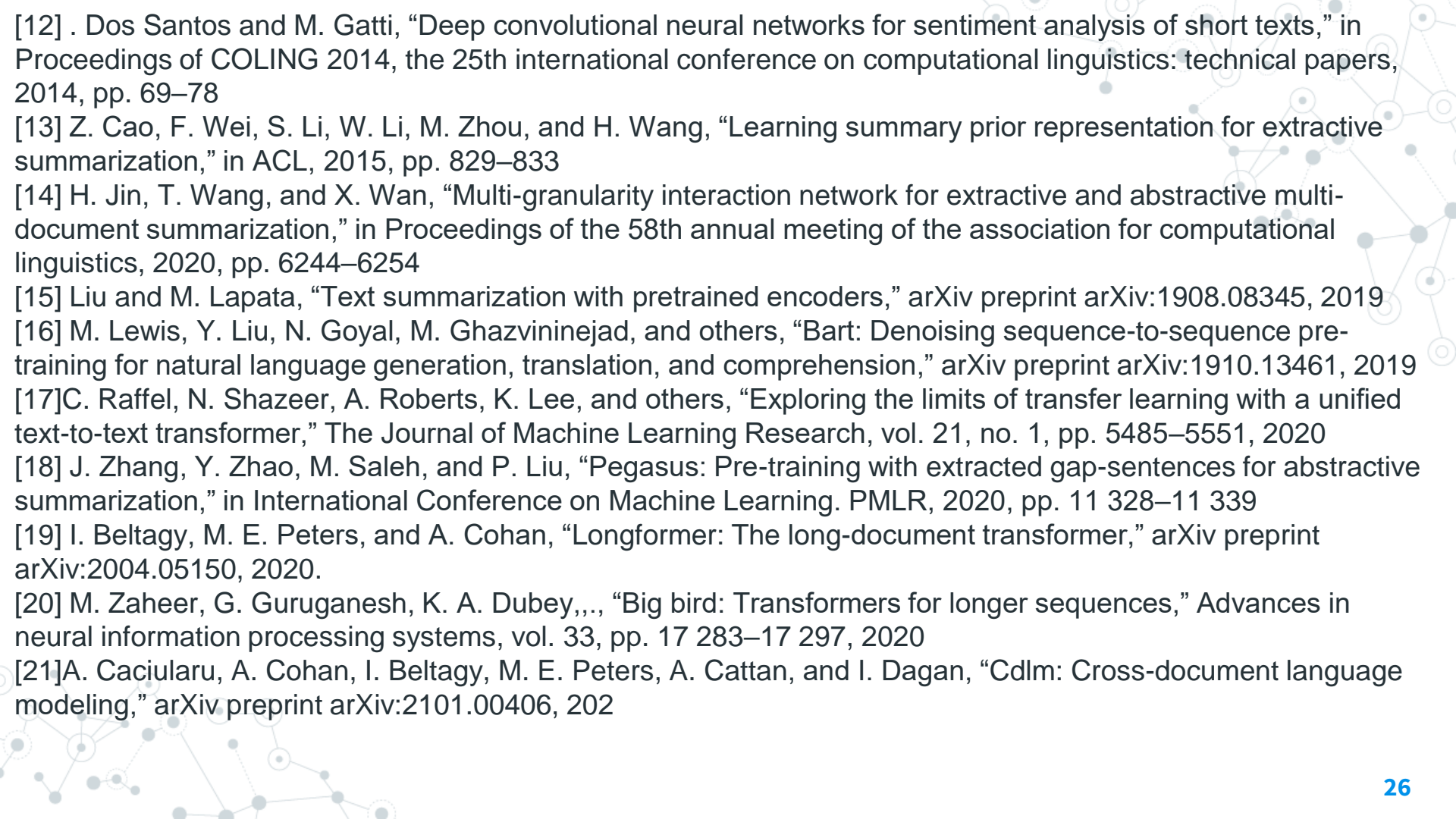
[30] S. Liu, J. Cao, R. Yang, and Z. Wen, "Generating a structured summary of numerous academic papers: Dataset and method," arXiv preprint arXiv:2302.04580, 2023

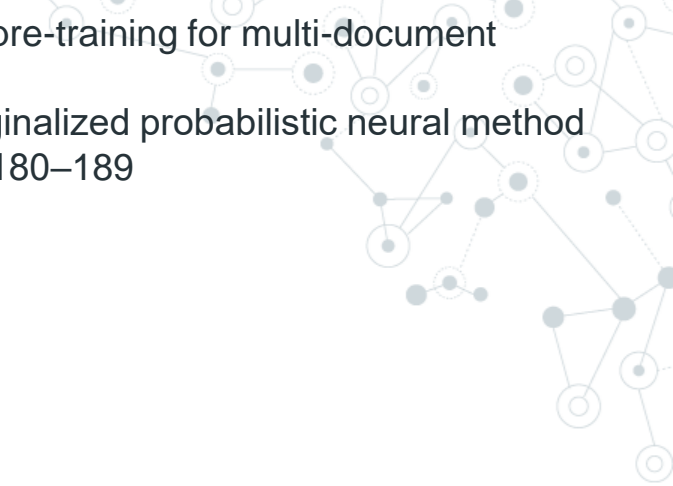

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by circles of varying sizes, some with solid centers and others with dashed outlines. The lines are thin and gray, creating a mesh-like structure.

REFERENCES



- 
- [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
- [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

A photograph of two parallel strings of clear, round light bulbs against a bright blue sky with soft, white clouds. The bulbs are slightly out of focus, creating a bokeh effect. The strings of lights run diagonally from the bottom left towards the top right.

THANK YOU...