

# Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles

**Yao Lu**  
Mila  
University of Waterloo  
yao.lu@uwaterloo.ca

**Yue Dong**  
Mila / McGill University  
yue.dong2  
@mail.mcgill.ca

**Laurent Charlin**  
Mila / HEC Montréal  
Canada CIFAR AI Chair  
lcharlin@gmail.com

## Abstract

Multi-document summarization is a challenging task for which there exists little large-scale datasets. We propose Multi-XScience, a large-scale multi-document summarization dataset created from scientific articles. Multi-XScience introduces a challenging multi-document summarization task: writing the related-work section of a paper based on its abstract and the articles it references. Our work is inspired by extreme summarization, a dataset construction protocol that favours abstractive modeling approaches. Descriptive statistics and empirical results—using several state-of-the-art models trained on the Multi-XScience dataset—reveal that Multi-XScience is well suited for abstractive models.<sup>1</sup>

## 1 Introduction

Single document summarization is the focus of most current summarization research thanks to the availability of large-scale single-document summarization datasets spanning multiple fields, including news (CNN/DailyMail (Hermann et al., 2015), NYT (Sandhaus, 2008), Newsroom (Grusky et al., 2018), XSum (Narayan et al., 2018a)), law (BigPatent (Sharma et al., 2019)), and even science (ArXiv and PubMed (Cohan et al., 2018)). These large-scale datasets are a necessity for modern data-hungry neural architectures (e.g. Transformers (Vaswani et al., 2017)) to shine at the summarization task. The versatility of available data has proven helpful in studying different types of summarization strategies as well as both extractive and abstractive models (Narayan et al., 2018a).

In contrast, research on the task of multi-document summarization (MDS) — a more general scenario with many downstream applications — has not progressed as much in part due to the

<sup>1</sup>Our dataset is available at <https://github.com/yaolu/Multi-XScience>

<b>Source 1 (Abstract of query paper)</b>
... we present an approach based on ... <b>lexical databases</b> and ... Our approach makes use of WordNet synonymy information to .... Incidentally, WordNet based approach performance is comparable with the training approach one.
<b>Source 2 (cite1 abstract)</b>
This paper presents a method for the resolution of lexical ambiguity of nouns ... The method relies on the use of the wide-coverage <b>noun taxonomy of WordNet and the notion of conceptual distance among concepts</b> ...
<b>Source 3 (cite2 abstract)</b>
Word groupings useful for language processing tasks are increasingly available ... This paper presents a method for <b>automatic sense disambiguation of nouns appearing within sets of related nouns</b> ... Disambiguation is performed with respect to WordNet senses ...
<b>Source 4 (cite3 abstract)</b>
In ... <b>word sense disambiguation...</b> integrates a <b>diverse set of knowledge sources ... including part of speech of neighboring words, morphological form</b> ...
<b>Summary (Related work of query paper)</b>
<b>Lexical databases</b> have been employed recently in <b>word sense disambiguation</b> . For example, ... [cite1] make use of a <b>semantic distance that takes into account structural factors in WordNet</b> ... Additionally, [cite2] combines the use of <b>WordNet and a text collection for a definition of a distance for disambiguating noun groupings</b> . ... [cite3] make use of <b>several sources of information ... (neighborhood, part of speech, morfological form, etc.)</b> ...

Table 1: An example from our Multi-XScience dataset showing the input documents and the related work of the target paper. Text is colored based on semantic similarity between sources and related work.

lack of large-scale datasets. There are only two available large-scale multi-document summarization datasets: Multi-News (Fabbri et al., 2019) and WikiSum (Liu et al., 2018). While large supervised neural network models already dominate the leaderboard associated with these datasets, obtaining better models requires domain-specific, high-quality, and large-scale datasets, especially ones for abstractive summarization methods.

We propose Multi-XScience, a large-scale dataset for multi-document summarization using *scientific articles*. We introduce a challenging multi-document summarization task: *write the re-*

lated work section of a paper using its abstract (source 1 in Tab. 1) and reference papers (additional sources).

Multi-XScience is inspired by the XSum dataset and can be seen as a multi-document version of extreme summarization (Narayan et al., 2018b). Similar to XSum, the “extremeness” makes our dataset more amenable to abstractive summarization strategies. Moreover, Table 4 shows that Multi-XScience contains fewer positional and extractive biases than previous MDS datasets. High positional and extractive biases can undesirably enable models to achieve high summarization scores by copying sentences from certain (fixed) positions, e.g. lead sentences in news summarization (Grenander et al., 2019; Narayan et al., 2018a). Empirical results show that our dataset is challenging and requires models having high-level of text abtractiveness.

## 2 Multi-XScience Dataset

We now describe the Multi-XScience dataset, including the data sources, data cleaning, and the processing procedures used to construct it. We also report descriptive statistics and an initial analysis which shows it is amenable to abstractive models.

### 2.1 Data Source

Our dataset is created by combining information from two sources: [arXiv.org](https://arxiv.org) and the Microsoft Academic Graph (MAG) (Sinha et al., 2015). We first obtain all arXiv papers, and then construct pairs of target summary and multi-reference documents using MAG.<sup>2</sup>

### 2.2 Dataset Creation

We construct the dataset with care to maximize its usefulness. The construction protocol includes: 1) cleaning the latex source of 1.3 millions arXiv papers, 2) aligning all of these papers and their references in MAG using numerous heuristics, 3) five cleaning iterations of the resulting data records interleaved with rounds of human verification.

Our dataset uses a query document’s abstract  $Q^a$  and the abstracts of articles it references  $R_1^a, \dots, R_n^a$ , where  $n$  is the number of reference articles cited by  $Q$  in its related-work section. The target is the query document’s related-work section segmented into paragraphs  $Q_1^{rw}, \dots, Q_k^{rw}$ , where  $k$

<sup>2</sup>Our dataset is processed based on the October 2019 dump of MAG and arXiv.

is the number of paragraphs in the related-work section of  $Q$ . We discuss these choices below. Table 1 contains an example from our dataset.

**Target summary:**  $Q_i^{rw}$  is a paragraph in the related-work section of  $Q$ . We only keep articles with an explicit related-work section as query documents. We made the choice of using paragraphs as targets rather than the whole related-work section for the following two reasons: 1) using the whole related work as targets make the dataset difficult to work on, because current techniques struggle with extremely long input and generation targets;<sup>3</sup> and 2) paragraphs in the related-work section often refer to (very) different research threads that can be divided into independent topics. Segmenting paragraphs creates a dataset with reasonable input/target length suitable for most existing models and common computational resources.

**Source:** the source in our dataset is a tuple  $(Q^a, R_1^a, \dots, R_n^a)$ . We only use the abstract of the query because the introduction section, for example, often overlaps with the related-work section. Using the introduction would then be closer to single-document-summarization. By only using the query abstract  $Q^a$  the dataset forces models to focus on leveraging the references. Furthermore, we approximate the reference documents using their abstract, as the full text of reference papers is often not available due to copyright restrictions.<sup>4</sup>

### 2.3 Dataset Statistics and Analysis

Dataset	# train/val/test	doc. len	summ. len	# refs
Multi-XScience	30,369/5,066/5,093	778.08	116.44	4.42
Multi-News	44,972/5,622/5,622	2,103.49	263.66	2.79
WikiSum	1, 5m/38k/38k	36,802.5	139.4	525

Table 2: Comparison of large-scale multi-document summarization datasets. We propose Multi-XScience. Average document length (“doc. len”) is calculated by concatenating all input sources (multiple reference documents).

In Table 2 we report the descriptive statistics of current large-scale multi-document summarization (MDS) datasets, including Multi-XScience. Compared to Multi-News, Multi-XScience has 60% more references, making it a better fit for the MDS settings. Despite our dataset being smaller than WikiSum, it is better suited to abstractive summarization as its reference summaries contain more

<sup>3</sup>10–20 references as input, 2–4 paragraphs as output

<sup>4</sup>Since our dataset relies on MAG for the reference paper as input, some reference papers are not available on arXiv. Our dataset contains all available paper information, including paper ids and corresponding MAG entry.

novel  $n$ -grams when compared to the source (Table 3). A dataset with a higher novel  $n$ -grams score has less extractive bias which should result in better abstraction for summarization models (Narayan et al., 2018a). Multi-XScience has one of the highest novel  $n$ -grams scores among existing large-scale datasets. This is expected since writing related works requires condensing complicated ideas into short summary paragraphs. The high level of abstractiveness makes our dataset challenging since models cannot simply copy sentences from the reference articles.

Datasets	% of novel $n$ -grams in target summary			
	unigrams	bigrams	trigrams	4-grams
CNN-DailyMail	17.00	53.91	71.98	80.29
NY Times	22.64	55.59	71.93	80.16
XSum	35.76	83.45	95.50	98.49
WikiSum	18.20	51.88	69.82	78.16
Multi-News	17.76	57.10	75.71	82.30
Multi-XScience	<b>42.33</b>	<b>81.75</b>	<b>94.57</b>	<b>97.62</b>

Table 3: The proportion of novel  $n$ -grams in the target reference summaries across different summarization datasets. The first and second block compare single-document and multi-document summarization datasets, respectively.

Datasets	LEAD			EXT-ORACLE		
	R-1	R-2	R-L	R-1	R-2	R-L
CNN-DailyMail	39.58	17.67	36.18	54.67	30.35	50.80
NY Times	31.85	15.86	23.75	52.08	31.59	46.72
XSum	16.30	1.61	11.95	29.79	8.81	22.65
WikiSum	38.22	16.85	26.89	44.40	22.59	41.28
Multi-News	43.08	14.27	38.97	49.06	21.54	44.27
Multi-XScience	<b>27.46</b>	<b>4.57</b>	<b>18.82</b>	<b>38.45</b>	<b>9.93</b>	<b>27.11</b>

Table 4: ROUGE scores for the LEAD and EXT-ORACLE baselines for different summarization datasets.

Table 4 reports the performance of the lead baseline<sup>5</sup> and the extractive oracle<sup>6</sup> for several summarization datasets. High ROUGE scores on the lead baseline indicate datasets with strong lead bias, which is typical of news summarization (Grenander et al., 2019). The extractive oracle performance indicates the level of “extractiveness” of each dataset. Highly-extractive datasets force abstractive models to copy input sentences to obtain a high summarization performance. Compared to the existing summarization datasets, Multi-XScience imposes much less position bias and requires a higher level

<sup>5</sup>The lead baseline selects the first- $K$  sentences from the source document as summary.

<sup>6</sup>The EXT-oracle summarizes by greedily selecting the sentences that maximize the ROUGE-L F1 scores as described in Nallapati et al. (2017).

of abstractiveness from models. Both results consolidate that Multi-XScience requires summarization models to “understand” source text (models cannot obtain a high score by learning positional cues) and is suitable for abstractive models (models cannot obtain a high score by copying sentences).

## 2.4 Human Evaluation on Dataset Quality

Two human judges evaluated the overlap between the sources and the target on 25 pairs randomly selected from the test set.<sup>7</sup> They scored each pair using the scale shown in Table 5.

Score	Criteria
4	75% - 100% facts (perfect coverage)
3	50% - 75% facts (major coverage)
2	25% - 50% facts (partial coverage)
1	less than 25% facts (poor coverage)

Table 5: Dataset quality evaluation criteria

The average human-evaluated quality score of Multi-XScience is  $2.82 \pm 0.4$  (95% C.I.). There is a large overlap between the reference abstracts and the targets’ related work based on this score<sup>8</sup> which highlights that the major facts are covered despite using only the abstract.

## 3 Experiments & Results

We study the performance of multiple state-of-the-art models using the Multi-XScience dataset. Detailed analyses of the generation quality are also provided, including quantitative and qualitative analysis in addition to the abstractiveness study.

### 3.1 Models

In addition to the *lead baseline* and *extractive oracle*, we also include two commonly used unsupervised extractive summarization models, *LexRank* (Erkan and Radev, 2004) and *TextRank* (Mihalcea and Tarau, 2004), as baselines.

For supervised abstractive models, we test state-of-the-art multi-document summarization models *HiMAP* (Fabbri et al., 2019) and *HierSumm* (Liu and Lapata, 2019a). Both deal with multi-documents using a *fusion* mechanism, which performs the transformation of the documents in the vector space. HiMAP adapts a pointer-generator

<sup>7</sup>We invited two PhD students who have extensive research experiences to conduct the dataset quality assessment on our scientific related-work summarization dataset.

<sup>8</sup>This is expected, as it is standard to discuss the key contribution(s) of a paper in its abstract.

model (See et al., 2017) with maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998; Lebanoff et al., 2018) to compute weights over multi-document inputs. HierSumm (Liu and Lapata, 2019a) uses a passage ranker that selects the most important document as the input to the hierarchical transformer-based generation model.

In addition, we apply existing state-of-the-art single-document summarization models, including *Pointer-Generator* (See et al., 2017), *BART* (Lewis et al., 2019) and *BertABS* (Liu and Lapata, 2019b), for the task of multi-document summarization by simply concatenating the input references. *Pointer-Generator* incorporates attention over source texts as a copy mechanism to aid the generation. *BART* is a sequence-to-sequence model with an encoder that is pre-trained with the denoising auto-encoder objective. *BertABS* uses a pretrained BERT (Devlin et al., 2019) as the encoder and trains a randomly initialized transformer decoder for abstractive summarization. We also report the performance of *BertABS* with an encoder (SciBert) pre-trained on scientific articles (Beltagy et al., 2019).

### 3.2 Implementation Details

All the models used in our paper are based on open-source code released by their authors. For all models, we use the default configuration (model size, optimizer learning rate, etc.) from the original implementation. During the decoding process, we use beam search (beam size=4) and tri-gram blocking as is standard for sequence-to-sequence models. We set the minimal generation length to 110 tokens given the dataset statistics. Similar to the CNN/Dailymail dataset, we adopt the anonymized setting of citation symbols for the evaluation. In our dataset, the target related work contains citation reference to specific papers with special symbols (e.g. cite<sub>2</sub>). We replace all of these symbols by a standard symbol (e.g. cite) for evaluation.

### 3.3 Result Analysis

**Automatic Evaluation** We report ROUGE Scores<sup>9</sup> and percentage of novel n-grams for different models on the Multi-XScience dataset in Tables 6 and 7. When comparing abstractive models to extractive ones, we first observe that almost all abstractive models outperform the unsupervised extractive models—TextRank and LexRank—by wide margins. In addition, almost all the abstractive models

<sup>9</sup>The scores are computed with ROUGE-1.5.5 script with option “-c 95 -r 1000 -n 2 -a -m”

significantly outperform the extractive oracle in terms of R-L. This further shows the suitability of Multi-XScience for abstractive summarization.

To our surprise, *Pointer-Generator* outperforms self-pretrained abstractive summarization models, such as *BART* and *BertABS*. Our analyses (Table 7) reveal that this model performs highly abstractive summaries on our dataset, indicating that the model chooses to generate rather than copy. *BART* is highly extractive with the lowest novel n-gram among all approaches. This result may be due to the domain shift of the self pre-training datasets (Wikipedia and BookCorpus) since the performance of *SciBertAbs* is much higher in terms of ROUGE-L. In addition, the large number of parameters in the transformer-based decoders require massive supervised domain-specific training data.

Models	ROUGE-1	ROUGE-2	ROUGE-L
Multi-doc Extractive			
LEAD	27.46	4.57	18.82
LEXRANK	30.19	5.53	26.19
TEXTRANK	31.51	5.83	26.58
EXT-ORACLE	38.45	9.93	27.11
Multi-doc Abstractive (Fusion)			
HIERSUMM(MULTI)	30.02	5.04	27.60
HIMAP(MULTI)	31.66	5.91	28.43
Multi-doc Abstractive (Concat)			
BERTABS	31.56	5.02	28.05
BART	32.83	6.36	26.61
SCIBERTABS	32.12	5.59	29.01
POINTER-GENERATOR	<b>34.11</b>	<b>6.76</b>	<b>30.63</b>

Table 6: ROUGE results on Multi-XScience test set.

Models	% of novel n-grams in generated summary			
	unigrams	bigrams	trigrams	4-grams
PG (CNNDM)	0.07	2.24	6.03	9.72
PG (XSUM)	27.40	73.33	90.43	96.04
PG	18.82	57.54	80.22	89.32
HIERSUMM	27.52	77.16	<b>95.03</b>	<b>98.51</b>
HIMAP	23.13	63.58	86.50	94.15
BART	8.15	30.13	44.53	51.75
BERTABS	<b>34.18</b>	<b>81.99</b>	95.70	98.64
SCIBERTABS	46.57	89.05	97.92	99.31

Table 7: The proportion of novel n-grams in generated summary. PG (CNNDM) and PG (XSUM) denotes the pointer-generator model performance reported by papers (See et al., 2017; Narayan et al., 2018b) trained on different datasets. All the remaining results are trained on Multi-XScience dataset.

**Human Evaluation** We conduct human evaluation on ext-oracle, HiMAP, and *Pointer-Generator*, since each outperforms others in their respective

<b>Groundtruth Related Work</b>
a study by @cite attempt to address the uncertainty estimation in the domain of crowd counting. this study proposed a scalable neural network framework with quantification of decomposed uncertainty using a bootstrap ensemble ... the proposed uncertainty quantification method provides additional auxiliary insight to the crowd counting model ...
<b>Generated Related Work (Oracle)</b>
in this work, we focus on uncertainty estimation in the domain of crowd counting. we propose a scalable neural network framework with quantification of decomposed uncertainty using a bootstrap ensemble. we demonstrate that the proposed uncertainty quantification method provides additional insight to the crowd counting problem ...
<b>Generated Related Work (HiMAP)</b>
in @cite, the authors propose a scalable neural network model based on gaussian filter and brute-force nearest neighbor search algorithm. the uncertainty of the uncertainty is used as a density map for the crowd counting problem. the authors of @cite proposed to use the uncertainty quantification to improve the uncertainty ...
<b>Generated Related Work (Pointer-Generator)</b>
our work is also related to the work of @cite, where the authors propose a scalable neural network framework for crowd counting. they propose a method for uncertainty estimation in the context of crowd counting, which can be seen as a generalization of the uncertainty ...

Table 8: Generation example of extractive oracle (EXT-ORACLE), HiMAP and Pointer-Generator (PG).

section of Table 6. For evaluation, we randomly select 25 samples and present the system outputs in randomized order to the human judges. Two human judges are asked to rank system outputs from 1 (worst) to 3 (best). Higher rank score means better generation quality. The average score is 1.54, 2.28 and 2.18 for ext-oracle, HiMAP, and Pointer-Generator, respectively. According to the feedback of human evaluators, the overall writing style of abstractive models are much better than extractive models, which provides further evidence of the abstractive nature of Multi-XScience.

In addition, we show some generation examples in Table 8. Since the extractive oracle is copied from the source text, the writing style fails to resemble the related work despite capturing the correct content. In contrast, all generation models can adhere to the related-work writing style and their summaries also the correct content.

## 4 Related Work

Scientific document summarization is a challenging task. Multiple models trained on small datasets exist for this task (Hu and Wan, 2014; Jaidka et al., 2013; Hoang and Kan, 2010), as there are no available large-scale datasets (before this paper). Attempts at creating scientific summarization datasets

have been emerging, but not to the scale required for training neural-based models. For example, CL-Scisumm (Jaidka et al., 2016) created datasets from the ACL Anthology with 30–50 articles; Yasunaga et al. and AbuRa’ed et al.<sup>10</sup> proposed human-annotated datasets with at most 1,000 article and summary pairs. We believe that the lack of large-scale datasets slowed down development of multi-document summarization methods, and we hope that our proposed dataset will change that.

## 5 Extensions of Multi-XScience

We focus on summarization from the text of multiple documents, but our dataset could also be used for other tasks including:

- Graph-based summarization: Since our dataset is aligned with MAG, we could use its graph information (e.g., the citation graph) in addition to the plain text as input.
- Unsupervised in-domain corpus: Scientific-document understanding may benefit from using using related work (in addition to other sources such as non-directly related reference manuals). It is worth exploring how to use unsupervised in-domain corpus (e.g., all papers from N-hop subgraph of MAG) for better performance on downstream tasks.

## 6 Conclusion

The lack of large-scale dataset has slowed the progress of multi-document summarization (MDS) research. We introduce Multi-XScience, a large-scale dataset for MDS using scientific articles. Multi-XScience is better suited to abstractive summarization than previous MDS datasets, since it requires summarization models to exhibit high text understanding and abstraction capabilities. Experimental results show that our dataset is amenable to abstractive summarization models and is challenging for current models.

## Acknowledgments

This work is supported by the Canadian Institute For Advanced Research (CIFAR) through its AI chair program and an IVADO fundamental research grant. We thank Daniel Tarlow for the original idea that lead to this work and Compute Canada for providing the computational resources.

<sup>10</sup>This is concurrent work.

## References

- Ahmed AbuRa'ed, Horacio Saggion, and Luis Chiruzzo. A multi-level annotated corpus of scientific papers for scientific document summarization and cross-document relation discovery.
- Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.
- Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.
- Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.
- Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6021–6026.
- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.
- Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 427–435.
- Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: an optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633.
- Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2016. Overview of the cl-scisumm 2016 shared task. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, pages 93–102.
- Kokil Jaidka, Christopher Khoo, and Jin-Cheon Na. 2013. Deconstructing human literature reviews—a framework for multi-document summarization. In *proceedings of the 14th European workshop on natural language generation*, pages 125–135.
- Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.
- Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081.
- Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.
- Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.
- Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018b. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.
- Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.
- Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.
- Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darin 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.