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
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.¹

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 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 leadboard 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-

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

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.
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 abstractiveness.

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 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.\(^2\)

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 \(R^a_1, \ldots, R^a_n\), 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 \(Q^w_1, \ldots, Q^w_k\), where \(k\) 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:** \(Q^w_i\) 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; 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 \((Q^a, R^a_1, \ldots, R^a_n)\). 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.\(^3\)

2.3 Dataset Statistics and Analysis

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

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

\(^{1}10–20\) references as input, \(2–4\) paragraphs as output

\(^{3}\)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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>% of novel n-grams in target summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unigrams</td>
</tr>
<tr>
<td>CNN-DailyMail</td>
<td>17.00</td>
</tr>
<tr>
<td>NY Times</td>
<td>22.64</td>
</tr>
<tr>
<td>XSum</td>
<td>35.76</td>
</tr>
<tr>
<td>WikiSum</td>
<td>18.20</td>
</tr>
<tr>
<td>Multi-News</td>
<td>17.76</td>
</tr>
<tr>
<td>Multi-XScience</td>
<td>42.33</td>
</tr>
</tbody>
</table>

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.

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

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. They scored each pair using the scale shown in Table 5.

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

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 andTarau, 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

---

5The lead baseline selects the first-K sentences from the source document as summary.
6The EXT-oracle summarizes by greedily selecting the sentences that maximize the ROUGE-L F1 scores as described in Nallapati et al. (2017).
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 denosing 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) pretrained 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. cite2). 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\(^9\) 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 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.

\(^9\)The scores are computed with ROUGE-1.5.5 script with option “-c 95 -r 1000 -n 2 -a -m”

3.3 Result Analysis

Automatic Evaluation

We report ROUGE Scores\(^9\) 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 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.

\(^9\)The scores are computed with ROUGE-1.5.5 script with option “-c 95 -r 1000 -n 2 -a -m”

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. cite2). 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\(^9\) 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 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.

\(^9\)The scores are computed with ROUGE-1.5.5 script with option “-c 95 -r 1000 -n 2 -a -m”

Human Evaluation

We conduct human evaluation on ext-oracle, HiMAP, and Pointer-Generator, since each outperforms others in their respective
Groundtruth Related Work

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 ...

Generated Related Work (Oracle)

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 ...

Generated Related Work (HiMAP)

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 ...

Generated Related Work (Pointer-Generator)

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. 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 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.

\(^{10}\) This is concurrent work.
References


References


