LTRC @MuP 2022: Multi-Perspective Scientific Document Summarization Using Pre-trained Generation Models

Ashok Urlana, Nirmal Surange, Manish Shrivastava
Language Technologies Research Center, KCIS
IIIT Hyderabad, India
{ashok.urlana,nirmal.surange}@research.iiit.ac.in,m.shrivastava@iiit.ac.in

Abstract

The MuP-2022 shared task focuses on multi-perspective scientific document summarization. Given a scientific document, with multiple reference summaries, our goal was to develop a model that can produce a generic summary covering as many aspects of the document as covered by all of its reference summaries. This paper describes our best official model, a fine-tuned BART$_{large}$, along with a discussion on the challenges of this task and some of our unofficial models including SOTA generation models. Our submitted model out performed the given, MuP 2022 shared task baselines on ROUGE-2, ROUGE-L and average ROUGE F1-scores. Code of our submission can be accessed here.

1 Introduction

With the rapidly growing research community, the volume of scientific papers being published every year is also going up. Which makes it nearly impossible for researchers to stay on top of the latest research. Scientific document summarization plays a crucial role in mitigating this problem. However, generating generic summaries for scientific documents is a non-trivial task due to their specific structure, varied content and inclusion of citation sentences. Scientific articles often represent salient information through tables, figures, and pseudo-codes (Altmami and Menai, 2020) and mathematical equations. And, generic text does not usually contain such elements.

The two widely used approaches for scientific document summarization are content-based (Collins et al., 2017; Nikolov et al., 2018) and citation-based (Nakov et al.; Abu-Jbara and Radev, 2011; Yasunaga et al., 2019). The former relies on traditional extractive and abstractive methods whereas, the latter locates the target paper by matching a portion of text with the citation sentences.

Almost all traditional summarization models, whether extractive or abstractive, follow supervised learning approach. That means, given a document the model learns to generate its summary based on its given gold (target) summary. However, in real world, summary writing is very subjective. For a given document, there could be multiple different yet valid summaries where each summary writer has written a summary of the same document from their perspective of the document. This subjectivity raises concerns about the evaluation ability of the model that is presented with only one gold summary. The MuP-2022 shared task is a novel attempt to address this concern. The goal of multi-perspective summarization task is to develop models that are capable of leveraging multiple gold summaries to generate one generic summary.

MuP-2022 shared task data contains a collection of scientific documents with multiple summaries. These summaries were collected by first taking (one or) multiple scientific peer reviews for each document and then extracting the introductory paragraph that summarizes the key contributions of the paper from the reviewer's perspective.

For this task, we explored several pretrained sequence-to-sequence models such as BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and ProphetNet (Qi et al., 2020). We also experimented with: a two-stage fine-tuning approach using the SciTLDR dataset (Cachola et al., 2020) and the divide and conquer approach, by (Gidiotis and Tsoumakas, 2020), that first divides the document into multiple sections to obtain section-wise summaries, and then aggregates all partial summaries to form the complete summary.

For the MuP 2022 shared task dataset, our fine-tuned BART$_{large}$ model remained the best among all our experiments by achieving 40.68 ROUGE-1 F1-score and 26.04 average ROUGE F1-score.

2 Related Work

Research on summarizing scientific documents has been widely explored in recent years. It is perti-

Proceedings of the Third Workshop on Scholarly Document Processing, pages 279–284
October 12–17, 2022.

	Train		Validation	
#Pairs	18934		3604	
#Unique Pairs	8382		1060	
	Text	Summary	Text	Summary
#Avg Words	2671.41	113.57	2671	115.13
#Avg Sentences	122.35	4.78	121.14	4.82

Table 1: MuP Data Statistics

nent to note that there is a great deal of variation in the density of information covered (Over and Yen, 2004), the level of details, and the organization of the content within the scientific document summaries. Recent work by (Fabbri et al., 2021) uses question threads from the Yahoo forum to build the multi-perspective answer summarization corpus. Meng et al., (2021) present FactSum that contains four summaries for each paper covers different aspects, they can provide summaries based on user requests.

A number of scholarly document summarization datasets, including PubMed and arXiv (Cohan et al., 2018), were used for training neural models ScisummNet (Yasunaga et al., 2019) and SciTLDR for extreme summarization (Cachola et al., 2020). Unlike these datasets, MuP2022 shared task organizers released a multi-perspective summarization dataset for scientific documents.

Various generation models, including BART, T5, ProphetNet, and PEGASUS, have shown great performance in summarization tasks. In particular, models like Big Bird (Zaheer et al., 2021) and Longformer (Beltagy et al., 2020) were released to handle long documents.

3 Corpus Description

The multi-perspective scientific document summarization task aims to generate a summary that covers various aspects of the document. Evaluating such a system with just one gold (or reference) summary negatively impacts the goal, as summaries are usually very subjective. Considering the fact that multiple summaries would help cover more different perspectives of the scientific document, which a single summary might have missed.

MuP2022 (Cohan et al., 2022) shared task data[1] contains multiple reference summaries for majority of the training set documents, and all of the development set documents also had a minimum of 3 reference summaries. The corpus consists of around 10K papers and 26.5K summaries. The average length of the summaries is 114.3 words long.

4 Methodology

Self-supervised pretrained models like BART (Lewis et al., 2020), T5 (Raffel et al., 2020), XLNet (Yang et al., 2019), ProphetNet (Qi et al., 2020), PEGASUS (Zhang et al., 2020) have been effective for many generative tasks. We experiment with these pre-trained models and fine-tune them on MuP dataset for this task.

BART is a transformer-based (Vaswani et al., 2017) standard sequence-to-sequence model modified to work as an auto-encoder (Lewis et al., 2020). A self-supervised autoencoder is trained on the corrupted text (addition of noise) and uses a language model to reconstruct the original text with the true replacement of corrupted tokens. BART uses five "noising" methods: token masking, token deletion, text infilling, sentence permutation, and document rotation.

T5 or **Text to Text Transfer Transformer** (Raffel et al., 2020) is a transformer-based approach that converts all the text-based language problems into the text-to-text format. This strategy allows the use of the same model architecture across a diverse set of tasks. T5 is pretrained on a multi-task mixture of supervised and unsupervised tasks using the common crawled corpus. We fine tune T5 base model on MuP corpus.

ProphetNet (Qi et al., 2020) is a sequence-to-sequence pretraining model. The unique objective of this model is to predict the future n-grams as the self-supervised training strategy. Unlike the traditional sequence-to-sequence models, ProphetNet is optimized by n-step ahead prediction instead of one-step-ahead prediction. We experimented with ProphetNet models with and without fine-tuned on the CNN/DailyMail dataset.

Utilizing SciTLDR The TLDR (Cachola et al., 2020) approach aims at creating extremely short summaries (TLDRs) for scientific documents. For this task, the authors introduced a SciTLDR dataset of 5400 TLDRs over 3200 papers.

DANCER (Gidiotis and Tsoumakas, 2020) Most of the extractive and abstractive methods for scientific document summarization typically consider the input as abstract and/or full text of the article to generate the abstract-like summary. In contrast, DANCER divides the source text into multiple sections, generates an individual summary for

[1] https://github.com/allenai/mup

Model	R-1	R-2	R-L	Avg R-f
Baseline	**40.8**	12.3	24.5	25.8
$BART_{large}$ cnn	40.68	**12.47**	**24.99**	**26.05**
DistilBART cnn	39.36	11.79	24.47	25.21
$BART_{base}$ cnn	39.12	11.42	23.8	24.78
$T5_{base}$	38.35	11.26	24.64	24.75
ProphetNet	38.15	11.45	24.25	24.62
$BART_{base}$	38.53	11.39	23.92	24.61
ProphetNet cnn	37.59	10.91	24.09	24.2
DANCER + BART	33.07	9.06	18.2	20.11
BART + Two-stage	32.51	6.82	20.64	19.99

Table 2: ROUGE scores for models fine-tuned on MuP2022 dataset

Parameters	BART	T5	ProphetNet
Max source length	1024	1024	512
Max target length	150	128	128
Min target length	56	30	56
Batch Size	1	1	1
Epochs	2	10	1
Vocab Size	50265	32128	30522
Beam Size	4	4	5
Learning Rate	5e-5	1e-4	5e-5

Table 3: Experimental Setup and Parameters Settings

each section, and aggregates the partial summaries to form the target summary.

5 Experiments

All of our experiments were performed on the same splits of train, validation and test sets as provided by the organizers. Table 1 shows the data statistics. We used NLTK tokenizer and the simplified version data released by the task organizers to report all the counts mentioned in Table 1.

The following subsections detail various categories of experiments. We hypothesise that various sections of the source document may contribute in multi-perspective reviews of the document reviews. The subsection 5.3 and 5.4 detail the experiments conducted, specifically, to capture various sections of the document.

5.1 Existing Pre-trained Generation models

We experimented with existing SOTA generation models like BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and ProphetNet (Qi et al., 2020). Table 3 details the general experimental setup for each.

Experiments were conducted with different versions of these models, such as DistilBART-cnn, $BART_{base}$, $BART_{base}$-cnn (base model of BART fine-tuned on CNN dataset), and ProphetNet-cnn.

Among all these, $BART_{large}$ achieved better performance for the MuP task. We use the $BART_{large}$ model fine-tuned on the CNN/DailyMail dataset (Hermann et al., 2015) to initialize our model.

5.2 Two Stage Fine-tuning

In order to follow TLDR (Cachola et al., 2020) approach, we attempted two stage fine tuning. Using the available checkpoints in the Hugging Face Transformers Library (Wolf et al., 2020), first we fine-tune the BART model on the SciTLDR dataset for 10 epochs with the max source and target token lengths of 1024 and 150 respectively. In the second stage, we fine-tune this model on the MuP dataset, with the same settings. However, as the bottom line of the Table 2 shows, this approach did not help with this MuP task.

5.3 Data Variation

The entire MuP dataset was released in two formats: one that consisted of the full-text of the scientific document along with meta-data and second, a simplified version of the source document. This simplified content is basically the pre-processed initial 2000 tokens of the documents' introduction sections.

We conducted a few experiments, with our submitted model, to investigate the contribution of various sections of these documents in the target summaries. For this, we created four categories of training, validation and test sets such that each category's source content consisted of one of the following combinations of sections of the source document:

1. ***Introduction*:** Only the introduction section of the document was used as the input to the BART model.

2. ***Abstract + Introduction*:** Both abstract and introduction sections, in concatenation, were utilized as the input for the BART model.

3. ***Abstract + Introduction + Conclusion*:** The BART model was fed with a combination of abstract, introduction and conclusion sections (if available) of the document.

4. ***Abstract + Conclusion*:** A combination of abstract and conclusion section was used as the input to the BART model.

First, we separately fine-tuned our $BART_{large}$ model using the training and validation data of

				Train & Val Data				Test Data			
R-1	R-2	R-L	Avg R-f	1	2	3	4	1	2	3	4
40.68	**12.47**	**24.99**	**26.05**	✓					✓		
40.67	12.5	24.93	26.03	✓						✓	
40.47	12.29	24.76	25.84			✓			✓		
40.34	12.28	24.79	25.8				✓				✓
40.33	12.28	24.75	25.79		✓				✓		
40.39	12.25	24.73	25.79			✓				✓	
40.23	12.32	24.77	25.77	✓							✓
40.23	12.17	24.6	25.67				✓		✓		
40.1	12.25	24.63	25.66		✓			✓			
40.22	12.13	24.54	25.63	✓				✓			

Table 4: Impact of Data Variations

each of these categories. Next, in each of these experiments all 4 models were tested with all 4 categories of test data. Table 4 shows the respective ROUGE f1-scores. Where, the checkmarks (✓) indicate the selected combination of training and test data category.

As shown in Table 4, the combination of '1' & '2' (i.e. only-introduction section for the training data and abstract + introduction for test data) outperforms all the rest. All these models were fine-tuned for two epochs and with the max source and target lengths of 1024 and 150, respectively.

5.4 Divide and Conquer Approach

Following the DANCER approach, we prepare the training, validation and test inputs by dividing each corresponding source documents into four sections: **Abstract, Introduction, Results and Discussion, and Conclusion**. We fine-tuned the BART model on each section of information separately and combined all the summaries at the end to get the final generated summary.

5.5 Impact of Hyperparameters

In order to find the optimal architecture for our BART$_{large}$ model, we experimented with number-of-epochs (1, 2, 3, 5) with default max-target-length of 128, where fine-tuning with 2 epochs showed better performance. We then tested for max-target-lengths (128, 150, 200) with 2 epochs. Where max-target-length 150 gave slightly better performance than the remaining. Tables 5 and 6 detail the corresponding ROUGE f1-scores.

Epochs	R-1	R-2	R-L	Avg R-f
1	40.5	12.48	24.88	25.95
2	**40.57**	**12.49**	**24.98**	**26.01**
3	40.31	12.23	24.8	25.78
5	40.35	12.02	24.59	25.65

Table 5: Impact of number-of-Epochs Variation

Epochs	Max Target Length	R-1	R-2	R-L	Avg R-f
2	128	40.57	**12.49**	24.98	26.01
	150	**40.68**	12.47	**24.99**	**26.05**
	200	40.67	12.47	24.99	26.04
5	128	**40.35**	12.02	24.59	25.65
	150	40.31	**12.1**	**24.66**	**25.69**

Table 6: Impact of Max-Target-Length Variation

6 Results & Discussion

For the MuP task, we experimented with various pre-trained generation models, a couple of scientific document summarization approaches, methods to cover different sections of the document and parameter settings. As shown in Table 2, among all of these the BART$_{large}$cnn (our submitted) model performed the best. This model was fine-tuned for 2 epochs with max-target-length 150 and data combination 1-2 (as mentioned in section 5.3). With this model, we secured 3rd rank in the MuP-2022 shared task.

While the MuP task considers summaries from multiple reviewers as different "perspectives", most of these summaries cover only the major contributions of the paper. These summaries, though diverse in their construction, do not look at the research paper from different points-of-view. We see a validation of this claim from the results in table 4, where model trained on "introduction" section alone outperforms all other combinations.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 500–509.

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2020. Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1804.05685*.

Arman Cohan, Guy Feigenblat, Tirthankar Ghosal, and Michal Shmueli-Scheuer. 2022. Overview of the first shared task on multi-perspective scientific document summarization (mup). In *Proceedings of the Third Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*.

Alexander R Fabbri, Xiaojian Wu, Srini Iyer, and Mona Diab. 2021. Multi-perspective abstractive answer summarization. *arXiv preprint arXiv:2104.08536*.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. *arXiv preprint arXiv:2106.00130*.

Preslav Nakov, Ariel Schwartz, and M Hearst. Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*.

Nikola I Nikolov, Michael Pfeiffer, and Richard HR Hahnloser. 2018. Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*.

Paul Over and James Yen. 2004. An introduction to duc-2004. *National Institute of Standards and Technology*.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.