# Team AINLPML @ MuP in SDP 2021: Scientific Document Summarization by End-to-End Extractive and Abstractive Approach

**Sandeep Kumar†, Guneet Singh Kohli*, Kartik Shinde†, Asif Ekbal†**
†Indian Institute of Technology Patna, India
∗Thapar Institute of Engineering and Technology, India
†(sandeep_2121cs29,kartik_1901ce16,asif)@iitp.ac.in
*guneetsk99@gmail.com

## Abstract

This paper introduces the proposed summarization system of the AINLPML team for the First Shared Task on Multi-Perspective Scientific Document Summarization at SDP 2022. We present a method to produce abstractive summaries of scientific documents. First, we perform an extractive summarization step to identify the essential part of the paper. The extraction step includes utilizing a contributing sentence identification model to determine the contributing sentences in selected sections and portions of the text. In the next step, the extracted relevant information is used to condition the transformer language model to generate an abstractive summary. In particular, we fine-tuned the pre-trained BART model on the extracted summary from the previous step. Our proposed model successfully outperformed the baseline provided by the organizers by a significant margin. Our approach achieves the best average Rouge F1 Score, Rouge-2 F1 Score, and Rouge-L F1 Score among all submissions.

## 1 Introduction

Automatic summarization involves distilling a document down to its essentials. There are two types of summarization techniques: abstractive summarization and extractive summarization. Abstractive summarization examines a document and creates a summary from it that may contain phrases that do not present in the original text. The more challenging goal is abstractive summarization, which is beneficial in fields like novels where phrases taken out of context are not a good foundation for producing a grammatical and cohesive summary. We are interested in summarizing scientific literature in this instance. Summarization of research papers can help in obtaining core ideas instantly and would help researchers all around the world in fastening the process of literature surveys.

It is well recognized that creating summaries of scientific papers is a difficult endeavour. The main question is why the article's abstract doesn't suffice since it summarizes the scientific article. Although an abstract has been written, there are many reasons for generating article summaries. First, one of the main problems with abstracts is that they do not include relevant information from the full text. Second, it presents the author's viewpoint on the unique characteristic in an incomplete and biased manner (Yang et al., 2016). Thirdly, no single summary meets all the user's needs (Reeve et al., 2007). In addition, the abstract does not cover all the impacts and contributions of the article (Elkiss et al., 2008) but rather what the author wishes to emphasize. As a result, the summary generated by such a system should be informative enough, cover all the critical sections of the input article, and provide the reader with essential information. Furthermore, (Yasunaga et al., 2019) discuss the impact factor of a scientific article. Summarization systems should accommodate the viewpoints of other researchers (i.e., citations) and the significant aspects highlighted by the article's authors in the abstract since the significance of papers may change over time.

Most existing summarizing research assumes only one best gold summary for each given material. Having just one gold summary limits our capacity to assess the effectiveness of summarizing algorithms because creating summaries is important to derive the significant aspects of any long document. Furthermore, because it takes subject matter experts a lot of time to read and comprehend lengthy scientific publications, annotating several gold summaries for scientific documents can be very expensive. The workshops aimed to promote the exploration of strategies for producing multi-perspective summaries. A novel summarizing corpus was provided that used information from peer-reviewed scientific articles to capture various viewpoints from the reader's perspective. In many different branches of science, peer reviews typi-



*Proceedings of the Third Workshop on Scholarly Document Processing*, pages 285–290
October 12–17, 2022.

cally begin with a paragraph that summarizes the most important contributions made by a work from the perspective of the reviewer, and each paper typically undergoes a number of different reviews.

This paper presents our approach to the MuP shared task(Cohan et al., 2022). We present an end-to-end approach to generate summaries of long scientific documents that uses the advantages of both extractive and abstractive approaches. Before producing a summary in an abstractive manner, we perform the extractive step, which is then used for conditioning the abstractor module. We first determined the section of a research paper. We took the Abstract, and the last few sentences of the introduction section as mostly authors summarize a few critical questions about the paper in these, such as, 'What is the contribution in the paper?', 'What is the novelty?', 'How is it different from previous works?'. From the rest of the portion of the document, we extracted the contributing sentences using a Large Language Model named ContriSci(Gupta et al., 2021). ContriSci is a BERT fine-tuned over sectional data from a research paper, capable of generating binary labels for a given sentence in that section which tells us if the sentence is contributing to the understanding of the section or not. After performing these extractive steps, we trained an abstractive model to form a final summary. Our experiments showed that jointly using extractive and abstractive models improves the summarization results.

## 2 Methodology

We propose an end-to-end pipeline approach to generate summaries automatically from scientific documents. Figure 1 shows an overview of our approach. We describe each component briefly as follows:

### 2.1 Extractive Model

The input to this model is the full text of the paper. Extractive Summarization deals with extracting pieces of text directly from the input document. Extractive Summarization can also be seen as a text classification task where we try to predict whether a given sentence will be part of the summary or not(Liu, 2019).

#### 2.1.1 Section Identification

Section information is essential as the reviewer often focuses on a few sections, such as the abstract and conclusion, more than other sections (Ghosh Roy et al., 2020). Section identification for any full scientific paper is not straightforward as there is no fixed pattern through which a template of a research paper is generalized. On close observation of the training data, We found that in training, only 60% of the data had a section named 'Conclusion' explicitly. Similarly, for 'Conclusion' similar problem was seen for generic sections such as 'Methodology' and 'Results'. Moreover, the section 'Conclusion' is not necessary the last section or the second last section of the paper. So, we found that the only sections uniformly available in each research paper were 'Introduction' and 'Abstract.'

#### 2.1.2 Contributing Sentence Identification

Apart from the 'Abstract' and last n[1] sentences of the introduction section, we also extract the contributing sentences using an attention-based deep neural model named ContriSci. ContriSci is a deep neural architecture that leverages Multi-task Learning to identify statements from a given research article that mention a contribution of the study. The model makes use of two auxiliary tasks: 1) Section Classification - classifying a given statement as belonging to a specific section of the paper, 2) Citation Classification - classifying whether a given statement consists of a citation within itself.

The authors generalize the specific sections of a conventional research paper into six categories - '*Title*', '*Abstract*', '*Introduction*', '*Background*', '*Method*', and '*Result*'. The study makes use of the NLPContributionGraph (NCG) data set (D'Souza et al., 2021) from Sem-Eval 2021 Task A [2]. The authors use set of predefined rules to annotate the dataset for the task of Section Classification. A specific research statement is fed into model together with the name of the section to which it belongs and the statements that surround it. Intuitively, this means that the model trains on more knowledge about the context in which a given research statement has been written. Given the peculiarities of the model and it's relevance to SDP, we choose to leverage it to enrich the extraction of textually salient statements.

[1] n=5. It was set empirically. We analyzed various values of n between 1 to 10 and chose the one that resulted in the best Rouge-1 F1 score

[2] https://ncg-task.github.io

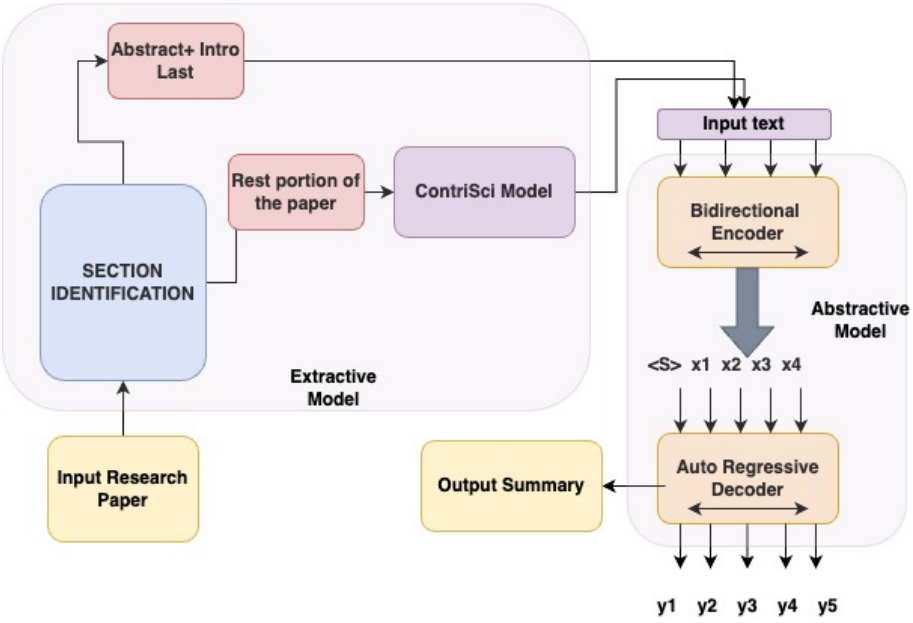


Figure 1: Architecture diagram of our proposed methodology.

| System | Rouge1 F | Rouge1 R | Rouge2 F | Rouge2 F | RougeL R | RougeL R | Avg Rouge F |
|---|---|---|---|---|---|---|---|
| Baseline | 40.80 | 44.20 | 12.33 | 13.50 | 24.48 | 26.81 | 25.87 |
| Other System | **41.36** | **43.29** | 12.52 | 13.20 | 24.83 | 26.21 | 26.24 |
| Our System | 41.08 | 42.96 | **13.29** | **13.98** | **25.36** | 26.62 | **26.58** |

Table 1: Experimental results of our model.(R:Recall, F:1 Score, Other System: Refers to the system with highest Rouge F1 in the leaderboard)

### 2.2 Abstractive Model

We use the BART autoencoder for pretraining sequence-to-sequence models. The structure of BART consists of two parts: an encoder and a decoder. The encoder part is a bidirectional encoder that corresponds to the structure of BERT (Vaswani et al., 2017), and the decoder part is an auto-regressive decoder following the settings of GPT. During the pretraining process, BART receives the corrupted document as input and performs the task of predicting the original uncorrupted document. In this way, BART can effectively learn contextual representations. When fine-tuned for the summarization task, the bidirectional encoder part encodes the original document, and the decoder part predicts the reference summary. BART obtains excellent performance on the summarization task. We gave the input to BART as follows:
Input text: Abstract $[SEP]$ INTRO_LAST $[SEP]$ Contributing sentences

Here the input to the BART model is Abstract, the last n sentences of the introduction(INTRO_LAST) and the contributing sentences separated by a token $[SEP]$. We use the BART fine-tuned on CNN/DailyMail dataset (Hermann et al., 2015) to initialize our model.

## 3 Experiments

In this section we discuss our results and analysis. The data set description (A) and experimental settings (B) can be found in the Appendix section.

### 3.1 Results and Discussion

In Table 1 the comparison of our best-submitted system has been made with the organizer's baseline model as well as the best performing system (based on Rouge1_f score (Lin, 2004)). Our methodology outperforms the baseline by a significant margin of 0.28 Rouge1_f score and 0.71 Avg Rouge F scores. Comparing our submission with the 'best leader board submission' shows that the submitted system performs well in Rouge2, RougeL, and overall avg Rouge F scores.

#### 3.1.1 Different inputs to the model

The submitted system had varied inputs passed through BART for summary generation. We report the result on the following combinations:

| Submitted Systems [Input to the BART while fine tuning] | Rouge1_F | Rouge2_F | RougeL_F | Avg Rouge F |
|---|---|---|---|---|
| Abstract + Full Paper | 40.53 | 12.02 | 24.32 | 25.62 |
| Abstract + Rule based selection from Intro | 40.62 | 12.22 | 24.22 | 25.74 |
| Abstract + Rule based selection from Full Paper | 40.78 | 12.19 | 24.27 | 25.79 |
| Abstract + Full Intro + ContriSci | 40.73 | 12.25 | 26.01 | 26.33 |
| Abstract + Intro Last + ContriSci | **41.08** | **13.29** | **25.36** | **26.58** |

Table 2: Ablation Study of our model.(F in the Rouge metrics refer to Rouge F1 Score)

- We performed the first set of experiments by tuning BART on Abstract + Full paper contents.
- Then we performed experiments by selecting contributing sentences from Abstract + Introductions of the paper and Abstract + Full paper. These contributing sentences were selected by defining rules to select sentences that contained words like 'propose,' 'demonstrate,' 'formulate,' 'contributes,' etc.
- The final set was formulating the approach of selecting contributing sentences using a ContriSciBERT(a pre-trained model used to identify whether a given sentence was a contributing sentence or not).

#### 3.1.2 Performance Analysis

We show the result of the experiments in Table 2. One of our significant experiments focused on exploiting sectional knowledge and selecting only sentences that concentrated on the substantial understanding of the paper. In particular, selecting contributing sentences helped to comprehend the paper's contribution. It assisted the subsequent model in generating a better-focused summary than other systems. Due to this we surpassed the baseline scores the organizers provided. In particular, we achieved an average Rouge F score of 26.58 when the Abstract + Intro Last + ContriSciBERT which is best among all the submissions made to the task. We also tested our result by passing the whole text of the introduction section as input. We achieved an avg Rouge F score of 26.33, which shows that it is better to give only the last portion of the introduction as it generally summarises the paper's contribution rather than proving the entire introduction to the subsequent summarization model. We also reported the result from extracting contributing sentences using generated rules. The result indicates that extracting contributing sentences from full papers is better than extracting them from only introduction section. We also report our system's scores on the Abstract + Full paper. The organizer used the same model as the baseline. The model produced a lower score than the baseline, perhaps because the organizers used better hyperparameters.

These analyses show the importance of the two-step approach to our proposed system. The first extractive summarization step ie: extracting the contributing sentences, the last part of the introduction section and the abstract written by the author assist the next abstractive step. It finally creates a focused summary highlighting the paper's contribution, motivation, etc of the paper. We perform a human evaluation of our summaries by hiring four human experts pursuing their masters in engineering and technology. They are well versed in NLP and machine learning. The ten summaries appear in entirely random order. We asked the responders to evaluate the summaries by rating them between 1 to 9 on the Likert Scale. The summaries generated by our model achieve the 7.5 Informativeness and 7 Coverage scores (described in Appendix Section C) compared to the golden summaries.

## 4 Conclusions

In this paper, we studied the Multi-Perspective Scientific Document Summarization task. We experimented with a joint model using extractive and abstractive approaches. The extractive approach supports the modelling of the document structure with a strong focus on which parts/sentences of a research paper to attend to while composing a summary, which significantly boosts the quality of the resultant output. On blind test corpora, our system ranks first wrt. to average Rouge F1 score. The results motivate towards experimenting with better extractive approaches in future which can improve the generation of abstractive summaries by feeding them ideal input data.

## 5 Acknowledgment

Sandeep Kumar acknowledges the Prime Minister Research Fellowship (PMRF) program of the Govt of India for its support.

## References


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.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. Semeval-2021 task 11: Nlpcontributiongraph–structuring scholarly nlp contributions for a research knowledge graph. *arXiv preprint arXiv:2106.07385*.

Aaron Elkiss, Siwei Shen, Anthony Fader, Günes Erkan, David J. States, and Dragomir R. Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *J. Assoc. Inf. Sci. Technol.*, 59(1):51–62.

Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta, and Vasudeva Varma. 2020. Summaformers @ LaySumm 20, LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 336–343, Online. Association for Computational Linguistics.

Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2021. Contrisci: A bert-based multitasking deep neural architecture to identify contribution statements from research papers. In *International Conference on Asian Digital Libraries*, pages 436–452. Springer.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

Lawrence H. Reeve, Hyoil Han, and Ari D. Brooks. 2007. The use of domain-specific concepts in biomedical text summarization. *Inf. Process. Manag.*, 43(6):1765–1776.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Shansong Yang, Weiming Lu, Zhanjiang Zhang, Baogang Wei, and Wenjia An. 2016. Amplifying scientific paper's abstract by leveraging data-weighted reconstruction. *Inf. Process. Manag.*, 52(4):698–719.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7386–7393. AAAI Press.


## A Data

The information from OpenReview[3], a platform for open and public publication of scientific research was provided. The corpus is composed of publications from venues including ICLR, NeurIPS, and AKBC. There are around 10,000 publications and 26.5 thousand summaries in the corpus (with an average number of 2.57 summaries per paper). Average word count for the summaries is 100.1. (space tokenized).

## B Experimental Settings

To train the ContriSci, we use an 80:10:10 split. We use the default hyperparameters with which ContriSci is trained. We use a learning rate of 1e-5 and an LR scheduler with Polynomial Decay and train the model for 5 epochs.

There are multiple summaries for a paper, so we have taken each paper's content and each summary as one instance to train the model[4]. We use a dynamic learning rate for the BART-based summarization, warm up 1000 iterations, and decay afterward. We set the batch size to 4. The gradient will accumulate every ten iterations, and we train all models for 6000 iterations on 1 GPU (NVIDIA A100 16GB). We save the best model with the highest Rouge1-F1 score based on the validation set. For the BART model, we use the implementation from the huggingface [5]. We use the BART large model pre-trained on CNN/DailyMail dataset.

## C Human Evaluation

We used the human evaluation as specified below :-

- Q1 (Readability): determines which of the summaries are most readable?
- Q2 (Informativeness): determines how much useful information about the reviews does the

[3] https://openreview.net/

[4] For example, if there are k summary of a paper, then we will create k instances of the paper.

[5] https://huggingface.co/

summary provide? You need to skim through
the original reviews to answer this.