Overview of the First Shared Task on Multi-Perspective Scientific Document Summarization (MuP)

Arman Cohan[1*] **Guy Feigenblat**[2] **Tirthankar Ghosal**[3] **Michal Shmueli-Scheuer**[4]

[1]Allen Institute for AI, Seattle, WA [2] Piiano Privacy Solutions
[3] ÚFAL, MFF, Charles University, CZ [4] IBM Research AI

armanc@allenai.org, guy@piiano.com
ghosal@ufal.mff.cuni.cz, shmueli@il.ibm.com

Abstract

We present the main findings of MuP 2022 shared task, the first shared task on multi-perspective scientific document summarization. The task provides a testbed representing challenges for summarization of scientific documents, and facilitates development of better models to leverage summaries generated from multiple perspectives. We received 139 total submissions from 9 teams. We evaluated submissions both by automated metrics (i.e., ROUGE) and human judgments on faithfulness, coverage, and readability which provided a more nuanced view of the differences between the systems. While we observe encouraging results from the participating teams, we conclude that there is still significant room left for improving summarization leveraging multiple references.[1]

1 Introduction

Generating summaries of scientific documents is known to be a challenging task as such documents are typically long and require domain expertise to fully comprehend them (Cohan et al., 2018; Cachola et al., 2020; Liu et al., 2022). The standard automated evaluation means in summarization compare system generated summaries with gold human written ones. At the same time, majority of existing work assumes only one single best gold summary for each given document. However, different readers of the same document can have different perspectives; therefore, there is often variability in human written summaries for a given document (Harman and Over, 2004). Having only one gold summary negatively impacts our ability to evaluate the quality of summarization systems through automated metrics (Harman and Over, 2004; Zechner, 1996). Also at training time this potentially prevents the model from capturing salient points with respect to different facets in the document (Hirsch et al., 2021). This is specially the case for longer documents where the summary compression ratio (ratio of length of the input document to the length of summary) is high (Cachola et al., 2020). While having multiple reference summaries for each document is desirable, human data collection can be expensive especially for long scientific documents.

To address this challenge, we introduce a new dataset and a new shared task to explore methods for generating multi-perspective summaries. We introduce a novel summarization corpus, MUP, leveraging data from scientific peer reviews to capture diverse perspectives from the reader's point of view. Our shared task similarly encourages development of methods to leverage multiple references. The dataset is collected from OpenReview,[2] an open publishing platform where peer reviews for some machine learning venues are publicly available. Peer reviews in various scientific fields often include an introductory paragraph that summarizes the main points and key contributions of a paper from the reviewer standpoint. For example, the first guideline to the reviewers in ACL review form[3] is to provide a "summary of the paper". In addition, each paper usually receives multiple reviews. Based on peer reviews, we collect a corpus of papers and their reviews from AI related venues such as ICLR, NeurIPS, and AKBC. We use carefully designed heuristics to only include first paragraphs of reviews that are summary-like. We manually check the summaries obtained from this approach on a subset of the data and ensure the high quality of the summaries. The corpus contains a total of 12K papers, and 27K summaries (with average number of 2.57 summaries per paper).

We next introduce MuP 2022, the first shared

[*] All authors contributed equally. Order is alphabetical.
[1]Our dataset is available at https://github.com/allenai/mup

[2]openreview.net
[3]https://aclrollingreview.org/reviewform

Proceedings of the Third Workshop on Scholarly Document Processing, pages 263–267
October 12–17, 2022.

task on multi-reference summarization with the goal of encouraging the community to develop better summarization methods for leveraging multiple references. Nine teams participated in the task with the top scoring models leveraging a range of transformer-based and graph-based models. Automated evaluation results show that while we observe notable progress in the task, there is ample room left for future improvements. We also conduct human evaluation on submitted systems and found out that while most system outputs are readable, they often struggle with the coverage aspect of summarization and they tend to miss some important information in the document.

2 Task

This section describes the MuP 2022 shared task.

2.1 Definition

The MuP task is basically an standard document summarization task where the goal is to generate a summary $\mathcal{S}_{gen}$ given a document $\mathcal{D}$, capturing its salient points. Teams were instructed to generate a summary for each of the papers in the MUP test set. The input is the full text content of papers along with section information. For each paper, the generated summary $\mathcal{S}_{gen}$ is evaluated against the set of m gold references $\langle \mathcal{S}_{g_1}, ..., \mathcal{S}_{g_m} \rangle$.

2.2 Evaluation and System Submissions

Following standard practice in summarization evaluation, we use ROUGE (Lin, 2004) as the primary evaluation metric. The average of the ROUGE-F scores obtained against the multiple summaries and averaged over ROUGE-1, ROUGE-2, and ROUGE-L was used for final ranking for the leaderboard. We used the unlimited length ROUGE version. In addition, we conducted human evaluation on a sample of summaries submitted by systems to get better insights about faithfulness, readability and coverage. The training set was released 50 days prior to the release of the hidden test set (papers content). Codalab framework[4] was used for the evaluation against the hidden test set. Participants were allowed to submit up to 25 submissions and the evaluation period lasted a month.

[4] https://codalab.lisn.upsaclay.fr/competitions/5676

3 Dataset Description

The MuP summarization dataset is collected using the publicly available peer review data, sidestepping the significant costs associated with manually creating multiple summaries for each scientific document.

3.1 Dataset Collection and Creation

We use the OpenReview API[5], to extract reviews from publicly open AI related venues such as ICLR, NeurIPS and AKBC. We extract fields including the paper title, summary (if exists, under the field "Summary") and the main review (under "Review" field). In addition, we use Science-Parse[6] to extract full text of the paper from the PDF. Science-Parse outputs a JSON record for each PDF, which among other fields, contains the title, abstract text, metadata (such as authors and year), and a list of the sections of the paper. Participants could leverage any type of additional metadata to improve their models.

After collecting the reviews we use parts of the review as a candidate summary for the paper as follows. Some conferences provide a review form that explicitly ask for a summary section ("Summary"). For example, starting from 2020 NeurIPS[7] asks the reviewers to "Summarize the paper motivation, key contributions and achievements in a paragraph". Similarly, in the ACL rolling review[8] reviewers are asked for a separate summary of the paper "Summary of the paper - Describe what this paper is about.". For those, we simply extract the summary section. When a summary field does not exist, we assume a common methodology that asks to describe what is the paper about, and what contributions does it make, followed by the main strengths and weaknesses. For example in ICLR 2021[9] reviewers were asked to "Summarize what the paper claims to contribute. List strong and weak points of the paper.". Here, we need to extract only the part that discusses the main contributions. We assume that the reviewers followed the review guidelines, and started with summarizing the main contributions, followed by a detailed description on

[5] https://openreview-py.readthedocs.io/

[6] https://github.com/allenai/science-parse

[7] https://nips.cc/Conferences/2020/PaperInformation/ReviewerGuidelines

[8] https://aclrollingreview.org/reviewform

[9] https://iclr.cc/Conferences/2021/ReviewerGuide#step-by-step

#Summaries	1	2	3	4	5	>5
#Papers	2276	3039	2867	1827	225	257

Table 1: Statistics of the MUP dataset.

the strengths and weaknesses. Thus, we extracted the first paragraph of the review section. To ensure that those paragraphs are indeed summaries and not opinions nor criticism (i.e., strengths and weaknesses), we followed Keith Norambuena et al. (2019), and used a lexicon-based approach to determine whether the paragraph carries a sentiment or not, in addition, we also removed paragraphs that contained individual pronouns (I, me, mine, myself). After these filtering process, two organizers of this task went through a random sample of 300 paragraphs, and annotated whether they are qualified as summaries. In total, 95% of the paragraphs were annotated as summaries. Table 1 summarizes the characteristics of the MUP dataset, which includes 10,491 summaries with an average length of 100.1 words long (space tokenized).

4 Systems

In this section, we overview the systems participating in the MuP shared task.

4.1 Baseline

As a simple baseline we use the BART-Large model (Lewis et al., 2020) further trained on CNN-DM summarization dataset (Hermann et al., 2015).[10] This baseline was made available to participants prior to the evaluation period.

4.2 Participant System Description

Although 18 teams registered, 9 teams participated (submitted their system runs). Here we briefly describe the approaches of the participating systems that provided us with a system description paper.[11]

Graph Attention Networks (GATS) (Akkasi, 2022) This work employs a Graph Attention Network-based extractive summarization approach for the task in hand. The approach is based on ranking the sentences in each of the discourse facets of the paper. Using Graph Attention Networks (GATs), the authors create a graph for each article after choosing three sentences that are closest to the ground truth summary. They define the rank of the sentences as the normalized average cosine similarity score between each sentence and the ground truth summaries. Since the ground truth summaries were not available for the test data, the authors use the sentences in the abstract as ground truth in the graph sentence selection and graph creation process.

GUIR (Sotudeh and Goharian, 2022) explored two different approaches to generate multi-perspective summaries. Their first approach learns a latent topic distribution using neural topic modeling (NTM) in the fine-tuning stage of a state-of-the-art abstractive summarizer (Longformer-Encoder-Decoder (Beltagy et al., 2020)), and the knowledge is shared between the topic modeling and text summarization task for summary generation. Their second approach involves adding a two-step summarizer that first extracts the salient sentences from the document and then writes abstractive summaries from those sentences. The second approach performs better on the official test set.

LTRC (Urlana et al., 2022) Their best-performing model is a fine-tuned BART-Large-CNN model which is same as the official baseline. They also experiment with several pre-trained sequence-to-sequence models (T5, ProphetNet, SciTLDR, DANCER) that first divides the document into multiple sections to obtain section-wise summaries, and then aggregates all partial summaries to form the complete summary. They experiment with different combination of paper-sections and found that only introduction section for the training and abstract + introduction for test data outperforms all the rest for the MuP task.

AINLPML (Kumar et al., 2022) This system adopts a two-stage approach for the task. In the first step, an extractive summarization step is used to identify the essential part of the paper. Their extraction step includes utilizing a contributing sentence identification model. In the next step, the authors finetune a BART model on the extracted summary generated from the previous step.

5 Results and Analysis

We only report the results of the teams who submitted their system papers in this section. Table 2 shows the comparative performance of the systems in MuP. The performance of the systems which

[10] We also tried training BART on scientific summarization datasets such as arxiv but did not achieve better results.

[11] Unfortunately, for system submissions without any report there is no way for us to know the details of the method and thus we exclude them from this overview paper.

Team	R-1	R-2	R-L	Avg
BART (baseline)	40.8	12.3	24.5	25.9
GATS Akkasi (2022)	33.7	7.4	17.7	19.6
LTRC (Urlana et al., 2022)	40.7	12.5	25.0	26.0
GUIR (Sotudeh and Goharian, 2022)	**41.4**	12.5	24.8	26.2
AINLPML (Kumar et al., 2022)	41.1	**13.3**	**25.4**	**26.6**

Table 2: Main results from the MuP 2022 shared task. R represents the ROUGE F1 metric.

used abstractive methods are generally better. In terms of average F_1 scores, team AINLPML (Kumar et al., 2022) produced the best performance, although GUIR and LTRC were pretty close. Except one, other teams were able to surpass the MuP baseline, although with small margins. Since the results of all systems were pretty close, in the next section we conduct human evaluation to gain better insights.

5.1 Human Evaluation

We asked domain experts in NLP (researchers with 10+ years experience in the field) to annotated a set of 20 randomly selected papers along with all system submissions for those papers. We asked the experts to rate the systems on a Likert scale (1-5), w.r.t three main qualities: faithfulness, readability, coverage, and Boolean rating for style ("review" vs. "summary"[12]). The experts could access the paper PDF and the ground-truth reviews. To evaluate faithfulness we asked them to first find important terms or phrases in the generated summary (e.g., datasets names, algorithms, etc), and then to look for them in the original paper and evaluate them in context. For readability, we asked annotators to take into account fluency, coherence and grammatical correctness. Finally, to understand coverage, annotators analyzed ground-truth summaries, and noticed that they tend to follow some structure, often a sentence or two for introduction, followed by methodology, and results. Hence, we expect to see such content covered in the generated summary. Similarly, if one important point is covered in one of the summaries but the generated summary fails to mention it, it gets penalized. Overall, the annotation task was time consuming with each annotator spending about 40 minutes on average per paper. Table 3 summarizes the average scores for the systems. Consistent with the automated evaluation results, AINLPML outperforms the rest of the systems in readability and coverage (and very close to leading also in faithfulness). From readability perspective, GATS received the lowest score, mainly due to low coherence. This is somewhat expected as their approach is extractive, and seems like no order was enforced (e.g., sometimes the introduction section appears last). Also since this approach is extractive, it achieves the highest faithfulness score. From the abstractive approaches, The BART baseline mainly suffered from the last sentence being trimmed in the middle. Further postprocessing/decoding methods could address these issues. LTRC summaries were much shorter than the other systems (on average 89 tokens vs. an average of 105 tokens of the rest of the systems), leading to generally lower coverage, but higher faithfulness. It is worth noting that the style in all the systems was annotated as "summary" - showing that the generated output looks like an actually summary than a peer review. Overall, while systems are able to get high performance in terms of faithfulness and readability, coverage remains a challenge and systems often tend to miss some important aspect of the paper.

Team	**Faithfulness**	**Readability**	**Coverage**
BART (baseline)	4.1	3.6	**3.9**
LTRC	4.4	4.6	3.6
GATS	5.0	2.7	2.4
GUIR	4.1	4.2	**3.9**
AINLPML	**4.4**	**4.7**	**3.9**

Table 3: Human evaluation (on a Likert scale 1-5).

6 Findings of MuP

Overall, our findings are summarized below:

- A general summarization baseline such as BART pretrained on news summarization dataset achieves decent results on the task.
- Combination of extractive and abstractive methods seem to work well for the task. This is inline

[12]To indicate that the generated output looks more like a peer review or like an actual summary.

with how human summarize longer documents by first identifying salient pieces of information and then aggregating this information.

- While we saw high scores in terms of faithfulness and readability, coverage remained a challenge.

- None of the participating systems focused on the multi-perspective aspect of the dataset. Submissions instead focused on general aspects of scientific document summarization such as length and specialized domain. This was somewhat unfortunate because our goal was to provide a testbed for developing methods for utilizing multiple summaries per document. We hope to see more of such models in future iterations of this task.

7 Conclusion and Future Directions

We present MuP, a new shared task and dataset of 27K summaries, which attracted attention from the community with 18 registered teams and 9 active submitting teams. Automated and human evaluation results suggest promising progress towards the task but we conclude that additional research is required, especially around utilization of multi references per document in the training process. For future iterations, we plan to extend the dataset by collecting reviews from additional venues. In addition, we plan to incorporate automatic measures of faithfulness as part of the leaderboard metrics.

References

Abbas Akkasi. 2022. Multi perspective scientific document summarization with graph attention networks (gats). In *Proceedings of the Third Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

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

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011*.

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*.

Donna Harman and Paul Over. 2004. The effects of human variation in duc summarization evaluation. In *Text Summarization Branches Out*, pages 10–17.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Eran Hirsch, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan. 2021. ifacetsum: Coreference-based interactive faceted summarization for multi-document exploration. *arXiv preprint arXiv:2109.11621*.

Brian Keith Norambuena, Exequiel Lettura, and Claudio Villegas. 2019. Sentiment analysis and opinion mining applied to scientific paper reviews. *Intelligent Data Analysis*, 23:191–214.

Sandeep Kumar, Guneet Singh, Kartik Shinde, and Asif Ekbal. 2022. Team ainlpml @ mup in sdp 2022: Scientific document summarization by end-to-end extractive and abstractive approach. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

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 *ACL*.

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.

Yixin Liu, Ansong Ni, Linyong Nan, Budhaditya Deb, Chenguang Zhu, Ahmed H Awadallah, and Dragomir Radev. 2022. Leveraging locality in abstractive text summarization. *arXiv preprint arXiv:2205.12476*.

Sajad Sotudeh and Nazli Goharian. 2022. Guir @ mup 2022: Towards generating topic-aware multi-perspective summaries for scientific documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

Ashok Urlana, Nirmal Surange, and Manish Shrivastava. 2022. Ltrc @mup 2022: Multi-perspective scientific document summarization using pre-trained generation models. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

Klaus Zechner. 1996. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.