Overview of the First Workshop on Scholarly Document Processing (SDP)

Muthu Kumar Chandrasekaran
Amazon, USA
cmkumar087@gmail.com

Guy Feigenblat
IBM Research AI, Israel
guyf@il.ibm.com

Dayne Freitag
SRI International, USA
daynefreitag@sri.com

Tirthankar Ghosal
Indian Institute of Technology Patna, India
tirthankar.pcs16@iitp.ac.in

Eduard Hovy
Carnegie Melon University, USA
hovy@cmu.edu

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences, Germany
philipp.mayr@gesis.org

Michal Shmueli-Scheuer
IBM Research AI, Israel
shmueli@il.ibm.com

Anita de Waard
Elsevier, USA
a.dewaard@elsevier.com

Abstract

Next to keeping up with the growing literature in their own and related fields, scholars increasingly also need to rebut pseudoscience and disinformation. To address these challenges, computational work on enhancing search, summarization, and analysis of scholarly documents has flourished. However, the various strands of research on scholarly document processing remain fragmented. To reach to the broader NLP and AI/ML community, pool distributed efforts and enable shared access to published research, we held the 1st Workshop on Scholarly Document Processing at EMNLP 2020 as a virtual event. The SDP workshop consisted of a research track (including a poster session), two invited talks and three Shared Tasks (CL-SciSumm, LaySumm and LongSumm), geared towards easier access to scientific methods and results. **Website**: https://ornlcda.github.io/SDProc

1 Workshop description

Over the past several years and at various venues, the Joint Workshop on Bibliometric-enhanced IR and NLP for Digital Libraries (**BIRNDL**[1]) (Cabanac et al., 2020; Mayr et al., 2018), the **CL-SciSumm** Shared Task (Chandrasekaran et al., 2019), and the International Workshop on Mining Scientific Publications (**WOSP**[2]) have established themselves as the principal venues for scholarly document processing (SDP) research. However, as these venues are collocated with conferences that are not focused on NLP, current solutions in this domain lag behind modern techniques generated by the greater NLP community.

The goal of SDP 2020 was to help foster cross-fertilization of ideas by bringing together people from different communities to leverage work on scientific literature and data. In doing so, we hope to create a premier meeting point to facilitate discussions converging towards solutions to open problems in SDP.

We believe that ACL events are the most appropriate venue for the SDP workshop for two reasons. First, ACL events are the premier venues for the confluence of NLP and ML and most of the cornerstone tasks in processing scholarly documents are NLP tasks. Improving machine understanding of scholarly semantics embedded in research papers is essential to further many tasks and applications in scholarly document processing. ACL events would, therefore, help integrate the broader NLP and AI/ML community with the distributed efforts in scholarly IR and Data Mining such that this field can progress as a more unified community. From our previous foray with IR and Data Mining we are convinced that delving into the language model of scholarly artefacts and improving machine understanding of scholarly semantics embedded in research papers is essential to further many tasks and applications in scholarly document processing. Second, we seek to bring together researchers and practitioners from various backgrounds focusing on different aspects of scholarly document processing. We believe the interdisciplinary nature of ACL venues would greatly assist in encouraging submissions from a diverse set of fields.

Topics. The topics of interest to SDP encompass all approaches to mining scholarly data and encourage submissions from all relevant communities, including:

[1] https://philippmayr.github.io/BIRNDL-WS/
[2] https://wosp.core.ac.uk/

Proceedings of the First Workshop on Scholarly Document Processing, pages 1–6

https://doi.org/10.18653/v1/P17

1. Information extraction, text mining and parsing of scholarly literature;
2. Reproducibility and peer review;
3. Lay Summarization (i.e., summaries created for non-experts) of individual and collections of scholarly documents;
4. Discourse modeling and argument mining;
5. Summarization and question-answering for scholarly documents;
6. Semantic and network-based indexing, search and navigation in structured text;
7. Graph analysis/mining including citation and co-authorship networks;
8. New scholarly language resources and evaluation;
9. Connecting and interlinking publications, data, tweets, blogs or their parts;
10. Disambiguation, metadata extraction, enrichment, and data quality assurance for scholarly documents;
11. Bibliometrics, scientometrics, and altmetrics approaches and applications;
12. Other aspects of scientific workflows including open access/science, and research assessment;
13. Infrastructures for accessing scientific publications and/or research data;
14. Results and research questions on the COVID-19 Open Research Dataset (CORD-19).

Workshop agenda. The SDP 2020 workshop[3] consisted of:

Two keynote talks, a Research Track (including a poster session) and a Shared Task Track with 3 separate shared tasks.

Keynotes. (1) **Kuansan Wang**, Managing Director, Microsoft Research Outreach Academic Services gave the first keynote titled: "Mitigating scholarly corpus biases with citations: A case study on CORD-19". *Abstract*: With the broad adoption of evidence based decision making processes, recent years have witnessed more frequent examples where biases in the datasets or the analytical algorithms lead to unfortunate and sometimes harmful outcomes. Being mindful of potential biases and actively taking measures to mitigate them have become a necessary second nature for scholars and decision makers alike. Citations in scholarly publications have long been known to represent the crowd-sourced collective judgments on scholarly communications and can be a valuable source of information in analyzing scholarly documents. This study describes a methodology that uses citations to identify biases in such corpus, using as an example the COVID-19 Open Research Dataset, or CORD-19, a corpus created to advance the development of intelligent technologies that can assist scientists in navigating through the voluminous literature of COVID-19. By expanding to articles in the citation networks seeded by CORD-19 with three distinct algorithms, it can be shown that CORD-19 has a strong tilt in favor of recent articles and uneven coverages in the topical fields and the publication venues. Using CORD-19 to identify critical knowledge and assess the journal importance, for example, will lead to different conclusions from the analyses based on the three expanded datasets, of which results largely agree with one another. CORD-19, however, does not appear to exhibit biases in describing research collaborations in terms of team sizes or geolocations. Currently, the three citation network traversal algorithms only utilize bibliographic records. How improvements can be made to them, such as through more sophisticated uses of citation contexts, will also be discussed.

(2) **Steinn Sigurðsson**, Scientific Director of arXiv, Professor in the Department of Astronomy & Astrophysics at The Pennsylvania State University gave the second keynote titled: "The future of arXiv and knowledge discovery in open science"[4]. *Abstract*: arXiv, the preprint server for the physical and mathematical sciences, is in its third decade of operation. As the flow of new, open access research increases inexorably, the challenges to keep up with and discover research content also become greater. I will discuss the status and future of arXiv, and possibilities and plans to make more effective use of the research database to enhance ongoing research efforts.

[3] The full program is available via https://ornlcda.github.io/SDProc/program.html.

[4] See the keynote paper in the SDP proceedings.

2 Research Track

In total, we received 34 papers for the research track. We accepted 9 papers for oral presentation (7 as full papers and 2 as short papers). We rejected 14 research paper submissions. In order to include a broader variety of contributions, we decided to invite all research papers with borderline scores as poster papers, leading us to accept 11 posters, which were presented in a separate virtual poster slot.

This year, the EMNLP organizers added a new resource, the "Findings of EMNLP"[5]. 520 EMNLP papers were accepted to Findings of EMNLP. Some of the authors of these Findings papers choose the SDP workshop as primary presentation venue. We accommodated and invited 3 "Findings" papers for oral presentation and 1 as a poster.

One demo paper was accepted as technical contribution in addition to the scientific program.

In the following, we list all contributions which were presented in some form at the workshop.

Oral presentations (full papers):

- Wu et al.: *Acknowledgement Entity Recognition in CORD-19 Papers*.
- Bhambhoria et al.: *A Smart System to Generate and Validate Question Answer Pairs for COVID-19 Literature*.
- Zhang et al.: *Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset*.
- Satish et al.: *The impact of preprint servers in the formation of novel ideas*.
- Berger et al.: *Effective Distributed Representations for Academic Expert Search*.
- Kim et al.: *Learning CNF Blocking for Large-scale Author Name Disambiguation*.
- Müller et al.: *Reconstructing Manual Information Extraction with DB-to-Document Backprojection: Experiments in the Life Science Domain*.

Oral presentations (short papers):

- Ling & Chen: *DeepPaperComposer: A Simple Solution for Training Data Preparation for Parsing Research Papers*.
- Medić & Snajder: *Improved Local Citation Recommendation Based on Context Enhanced with Global Information*.

Poster presentations:

- Ozyurt: *On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining*.
- Kashyap & Kan: *SciWING – A Software Toolkit for Scientific Document Processing*.
- Li et al.: *Multi-task Peer-Review Score Prediction*.
- Basu et al.: *ERLKG: Entity Representation Learning and Knowledge Graph based association analysis of COVID-19 through mining of unstructured biomedical corpora*.
- Asakura et al.: *Towards Grounding of Formulae*.
- van Dongen et al.: *SChuBERT: Scholarly Document Chunks with BERT-encoding boost Citation Count Prediction*.
- de Buy Wenniger et al.: *Structure-Tags Improve Text Classification for Scholarly Document Quality Prediction*.
- Ding et al.: *Cydex: Neural Search Infrastructure for the Scholarly Literature*.
- Patel et al.: *On the Use ofWeb Search to Improve Scientific Collections*.
- Goldfarb-Tarrant et al.: *Scaling Systematic Literature Reviews with Machine Learning Pipelines*.
- Kang et al.: *Document-Level Definition Detection in Scholarly Documents: Existing Models, Error Analyses, and Future Directions*.

Poster (demo paper):

- Fadaee et al.[6]: *A New Neural Search and Insights Platform for Navigating and Organizing AI Research*.

[5] https://2020.emnlp.org/blog/2020-04-19-findings-of-emnlp

[6] This paper was accepted as technical demo beside the research program.

EMNLP 2020 Findings papers: Cao et al., Noh & Kavuluru, Subramanian et al. were presented as short papers; Kobs et al. was presented as a poster.

- Cao et al.: *Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora*.
- Noh & Kavuluru: *Literature Retrieval for Precision Medicine with Neural Matching and Faceted Summarization*.
- Subramanian et al.: *MedICaT: A Dataset of Medical Images, Captions, and Textual References*.
- Kobs et al.: *Where to Submit? Helping Researchers to Choose the Right Venue*.

3 Shared Task Track

In addition to the research track, SDP hosted three Shared Tasks. Details of the task, results and overview are provided in a companion paper, '*Overview and Insights from the First Workshop on from Scholarly Document Processing: Shared Tasks: CL-SciSumm, LaySumm and LongSumm*' (Chandrasekaran et al., Forthcoming). We added these since summarization is an important and challenging effort within scholarly document processing, as the number and complexity of scientific papers increases exponentially, and making them accessible to both a lay and professional audience becomes increasingly important.

3.1 CL-SciSumm

CL-SciSumm is the first medium-scale shared task on scientific document summarization in the computational linguistics domain with over 500 documents annotated for their citation and citation targets and over a 1000 more documents with human annotated summaries inherited and integrated from SciSummNet (Yasunaga et al., 2019). Last year's CL-SciSumm shared task introduced large scale training datasets, both annotated from ScisummNet and auto-annotated. For this year's task, systems were provided with a Reference Paper (RP) and 10 or more Citing Papers (CPs) that all contain citations to the RP, which they used to summarise the RP. This was evaluated against abstract, citation-based summaries and human-written summaries with ROUGE. The shared task attracted 50+ registrations and 11 final system submissions. Importantly, we have now released the gold standard labels for our hitherto blind test set[7] which can serve as a public benchmark for evaluations on the CL-SciSumm corpus.

3.2 LaySumm

The **LaySumm** summarization task considers automating the generation of a Lay Summary: a text of about 70–100 words intended for a non-technical audience that explains, succinctly and without using technical jargon, the overall scope, goal, and potential impact expressed in a scientific paper. The corpus for this task comprised 572 full-text papers with lay summaries, in a variety of domains, including archaeology, hematology, and engineering, made available by *Elsevier*

The Lay summaries had to be representative of the content, comprehensible, and interesting to a lay audience. The intrinsic evaluation was done by ROUGE, through the CodaLabs Platform[8] In addition, a subset of randomly selected summaries underwent human evaluation by a team of science journalists and communicators for comprehensiveness, legibility, and interest. Authors were also asked to provide an automatically generated lay summary of their own paper together with their contribution.

3.3 LongSumm

The **LongSumm** task aims at creating long summaries of around 600 words. Often, for researchers, short summaries (e.g., abstract) are not detailed enough. Thus, longer summaries are mainly intended for helping researchers understand the gist of a paper without the need to read it entirely. The corpus for this task includes a training set that consists of 1705 extractive summaries, and 531 abstractive summaries of NLP and Machine Learning scientific papers. The extractive summaries are based on video talks from associated conferences (Lev et al., 2019) while the abstractive summaries are based on blog posts created by NLP and ML researchers. The test set consists of 22 abstractive summaries for evaluating the submissions. In total, 9 systems participated in the task, with a total of 100 submissions. The

[7] https://github.com/WING-NUS/scisumm-corpus/tree/master/data/Test-Set-2018-Gold

[8] https://competitions.codalab.org/competitions/25516#learn_the_details

evaluation was conducted using the ROUGE measure (Lin, 2004) and executed on a public leaderboard[9]. In addition, a subset of randomly selected summaries, of the top ranked systems, was evaluated by experts

4 Workshop Overview and Outlook

The organizers were gratified by both the size and breadth of the response to the inaugural edition of SDP. The subjects of accepted papers and posters ranged from end uses of the scholarly literature (such as search, recommendation, or literature curation) to challenges associated with automated understanding (such as entity recognition and disambiguation or formula grounding), to adaptations of recent successes in the broader field of NLP (such as generation or question answering). It is apparent that automated processing of the scholarly literature is a problem that meets with substantial interest. And it seems likely that we are observing the beginnings of a research community with a narrow enough focus to make rapid progress, but a broad enough set of concerns to offer ample opportunities for cross-pollination.

To a first approximation, we regard SDP as a confluence of three communities: NLP, information retrieval, and scientometrics. Given our colocation with EMNLP, it is perhaps not surprising that the majority of our submissions emphasized NLP. Certainly, our shared tasks all share a research focus, summarization, that is a traditional NLP problem area. As we consider future iterations of the workshop, we are discussing ways to increase its subject diversity. We have begun by identifying a more varied set of shared tasks, each highlighting challenges unique to the automated processing of the scholarly literature. As we proceed with planning and advertising, a key objective will be to elicit high-quality submissions from researchers interested in the uses and metalinguistic aspects of scholarly communication.

5 Organising and Steering Committees

A formal Organizing Committee and a Steering Committee helped guide the successful organisation first SDP. We thank all members for their help in reviewing all submissions, submitting the workshop proposal, and planning the final program: C. Lee Giles, Pennsylvania State University, USA; Min-Yen Kan, National University of Singapore; Petr Knoth, Open University, UK; Robert Patton, Oak Ridge National Laboratory, USA; Dragomir Radev, Yale University, USA; Jie Tang, Tsinghua University, China; Kuansan Wang, MSR Outreach Academic Services, USA; Bonnie Webber, University of Edinburgh, UK.

[9] https://aieval.draco.res.ibm.com/challenge/39/

6 Conclusion

The scholarly literature has long served as a rich source of interesting and challenging problems for computer science. Recent events regarding misinterpretation of scholarly information accentuate the importance of better approaches to the automated processing of scholarly literature.

We hope that this event helps to connect these challenges to use cases, fostering solutions that ultimately improve the practice of scholarship and serve society.

References

Guillaume Cabanac, Ingo Frommholz, and Philipp Mayr. 2020. Bibliometric-Enhanced Information Retrieval 10th Anniversary Workshop Edition. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, volume 12036, pages 641–647. Springer International Publishing, Cham.

M. K. Chandrasekaran, G. Feigenblat, Hovy. E., A. Ravichander, M. Shmueli-Scheuer, and A De Waard. Forthcoming. Overview and insights from scientific document summarization shared tasks 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing (SDP 2020)*.

Muthu Kumar Chandrasekaran, Michihiro Yasunaga, Dragomir Radev, Dayne Freitag, and Min-Yen Kan. 2019. Overview and results: Cl-scisumm shared task 2019. *arXiv preprint arXiv:1907.09854*.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. Talksumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. *arXiv preprint arXiv:1906.01351*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Philipp Mayr, Ingo Frommholz, Guillaume Cabanac, Muthu Kumar Chandrasekaran, Kokil Jaidka, Min-Yen Kan, and Dietmar Wolfram. 2018. Introduction

to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). *International Journal on Digital Libraries*, 19(2-3):107–111.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri Irene Li Dan, and Friedman Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks.