Overview of the Third Workshop on Scholarly Document Processing

Arman Cohan[a] **Guy Feigenblat**[b] **Dayne Freitag**[c]
Tirthankar Ghosal[d] **Drahomira Herrmannova**[e] **Petr Knoth**[f]
Kyle Lo[a] **Philipp Mayr**[g] **Michal Shmueli-Scheuer**[h]
Anita de Waard[e] **Lucy Lu Wang**[a,i]

Abstract

With the ever-increasing pace of research and high volume of scholarly communication, scholars face a daunting task. Not only must they keep up with the growing literature in their own and related fields, scholars increasingly also need to rebut pseudo-science and disinformation. These needs have motivated an increasing focus on computational methods for enhancing search, summarization, and analysis of scholarly documents. However, the various strands of research on scholarly document processing remain fragmented. To reach out to the broader NLP and AI/ML community, pool distributed efforts in this area, and enable shared access to published research, we held the 3rd Workshop on Scholarly Document Processing (SDP) at COLING as a hybrid event (https://sdproc.org/2022/). The SDP workshop consisted of a research track, three invited talks and five Shared Tasks: 1) MSLR22: Multi-Document Summarization for Literature Reviews, 2) DAGPap22: Detecting automatically generated scientific papers, 3) SV-Ident 2022: Survey Variable Identification in Social Science Publications, 4) SKGG: Scholarly Knowledge Graph Generation, 5) MuP 2022: Multi Perspective Scientific Document Summarization. The program was geared towards NLP, information retrieval, and data mining for scholarly documents, with an emphasis on identifying and providing solutions to open challenges.

[a]Allen Institute for AI, USA
[b]Piiano Privacy Solutions
[c]SRI International, USA
[d]ÚFAL, MFF, Charles University, Czech Republic
[e]Elsevier, USA
[f]The Open University, UK
[g]GESIS –– Leibniz Institute for the Social Sciences, Germany
[h]IBM Research AI, Haifa Research Lab, Israel
[i]University of Washington, USA

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, and the International Workshop on Mining Scientific Publications (**WOSP**[2]) (Knoth et al., 2020) have established themselves as the principal venues for research in scholarly document processing (SDP). 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.

In 2020, the first **SciNLP** workshop[3] was held online at the AKBC 2020 conference; the workshop brought together interested parties in a talk series focused on various aspects of scientific NLP. The first **Scholarly Document Processing** (SDP) workshop then took place in co-location with the EMNLP 2020 conference as an online workshop (see overview in Chandrasekaran et al. (2020)), and provided a dedicated venue for those working on SDP to submit and discuss their research. Following these successes and the clear appetite for venues to foster discussions around scholarly NLP, SDP 2021 co-located at NAACL, again aimed to connect researchers and practitioners from different communities working with scientific literature and data and created a premier meeting point to facilitate discussions on open problems in SDP.

Program The SDP 2022 workshop consisted of three Keynote talks, a Research Track and a Shared Task Track. The full program with links to papers, videos and posters is available at

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

Proceedings of the Third Workshop on Scholarly Document Processing, pages 1–6
October 12–17, 2022.

https://sdproc.org/2022/program.html.

2 Keynotes

This year, we had 3 keynote speakers discussing a variety of recent advancements in scholarly document processing: Min-Yen Kan (National University of Singapore), Sophia Ananiadou (University of Manchester), and Andrew Head (University of Pennsylvania). More talk info provided below:

Title "Scholarly Document Processing Research in the Age of AIs".

Speaker Min-Yen Kan

Abstract Artificial Intelligence is poised to impact many fields, but how will the rise of AI impact the way that we do science and scholarly work? Thomas Kuhn, in his philosophical analyses of sciences coined the term "paradigm shift" to describe the resultant progress in science theory when the normal science of an existing paradigm collides with theory-unaccountable, replicable observations. With scientists in AI still expecting key discoveries to be made, will we expect a new paradigm to overturn current normal science in AI and other fields? Will the age of accelerations, as defined by Thomas Friedman, hold sway over how real-world contexts are either accounted for or discarded by research practitioners and scholars alike? I relate my perspective on how normal science and paradigm shifting science relate to the notion of research, fast and slow, and how scholarly document processing can facilitate the mean and variance in science discovery. I give an opinionated view of the importance of scholarly document processing, as a meta-research agenda that can either aid thoughtful slow research, or be leveraged to further exacerbate acceleration of normal science.

Title "Biomedical Text Summarisation: Methods and Challenges"

Speaker Sophia Ananiadou

Abstract Biomedical text summarization techniques are used to support users in accessing information efficiently, by retaining only the most important semantic information contained within documents. Text summarization is important in a variety of scenarios, including systematic reviews (synthesis), evidence-based medicine, clinical decision support, etc. I will discuss current trends in biomedical text summarization, the use of pre-trained language models (PLMs), benchmarks, evaluation measures and challenges faced in both extractive and abstractive methods. In particular, I will examine how to extract salient sentences by exploiting both local and global contexts and explore how the integration of fine-grained medical knowledge into PLMs can improve extractive summarisation.

Title "Exploring How Intelligent Interfaces Can Support the Reading of Scholarly Articles"

Speaker Andrew Head

Abstract In this talk, I share a vision of interactive research papers, where user interfaces surface information for readers when and where they need it. Grounded in tools that I and my collaborators have developed, I discuss what it takes to design reading interfaces that (1) surface definitions of terms where readers need them (2) explain the meaning of math notation and (3) convey the meaning of jargon-dense passages in simpler terms. In our research, we have found that effective reading support requires not only sufficient document processing techniques, but also the careful presentation of derived information atop visually complex documents. I discuss tensions and solutions in designing interactive papers, and identify future research directions that can bring about powerful augmenting reading experiences.

3 Research Track

We invited submissions from all communities demonstrating usage of and challenges associated with natural language processing, information retrieval, and data mining of scholarly and scientific documents. Relevant topics included:

1. Representation learning
2. Information extraction
3. Summarization
4. Generation
5. Question answering
6. Discourse and argumentation mining
7. Network analysis
8. Bibliometrics, scientometrics, and altmetrics
9. Reproducibility
10. Peer review
11. Search and indexing
12. Datasets and resources
13. Document parsing

14. Text mining
15. Research infrastructure, and others.

In total, we accepted 18 submissions for the research track for presentation.

4 Shared Task Track

SDP 2022 hosted five shared tasks. Each shared task had its own organizing committee consisting of several members of the SDP 2022 organizers and/or other collaborators. Shared task presentations were held online in parallel sessions to the main SDP workshop. See short descriptions of the shared tasks below. Detailed overview papers of the shared tasks are referred to and followed in the proceedings.

4.1 Multi-document Summarization for Systematic Reviews (MSLR2022)

Organizers: Lucy Lu Wang, Jay DeYoung, and Byron Wallace

Systematic literature reviews aim to comprehensively summarize evidence from all available studies relevant to a question, and provide the highest quality evidence towards clinical care. Reviews are expensive to produce manually and quickly go out of date (Shojania et al., 2007); (semi-)automation via NLP may facilitate faster evidence synthesis without sacrificing rigor. Toward this end, we provided two datasets of reviews and studies derived from the scientific literature to study the task of generating review summaries (DeYoung et al., 2021; Wallace et al., 2020). We also encouraged submissions extending our task/datasets, e.g., proposing scaffolding tasks, methods for model interpretability, and improved automated evaluation methods. We received submissions from 6 teams, with a total of 10 public submissions to the Cochrane and MSˆ2 subtask leaderboards. We observed modest improvements in task performance as assessed by automated evaluation metrics, and gained significant insights into the remaining challenges for this task. Systems reports submitted by 5 teams are included in the workshop proceedings along with an overview paper (Wang et al., 2022) summarizing potential directions for future work.

4.2 Detecting automatically generated scientific papers (DAGPap22)

Organizers: Yury Kashnitsky, Drahomira Herrmannova, Anita de Waard, Georgios Tsatsaronis, Catriona Fennell, and Cyril Labbé

Can we automatically distinguish machine-generated papers from those written by humans? For this challenge, we provided a corpus of over 4,000 papers that are (probably) synthetic to some extent, based on the work of Cabanac et al. (2021), as well as documents collected by our publishing and editorial teams. As a control, we provided a corpus of open access human-written papers from the same scientific domains. We also encouraged contributions that extended this dataset with other computer-generated scientific papers, or papers that propose valid metrics to assess automatically generated papers against those written by humans. The DAGPap22 overview paper is available at Kashnitsky et al. (2022).

4.3 Survey Variable Identification in Social Science Publications (SV-Ident 2022)

Organizers: Tornike Tsereteli, Yavuz Selim Kartal, Simone Paolo Ponzetto, Andrea Zielinski, Kai Eckert, Philipp Mayr

The **SV-Ident 2022**[4] task is the first shared task on survey variable identification in the Social Science domain. Social Science literature often uses and references survey datasets, which contain sometimes hundreds of items or questions, called *survey variables* or *variables*. Studies may focus on and reference only a specific subset of these variables. While survey datasets that are used in a publication are typically referenced explicitly in-text using a bibliographic citation, individual variables are often only referenced ambiguously. This lack of explicit linking limits access to research along the FAIR principles.

The dataset for SV-Ident contains 5,972 expert-annotated sentences (with and without variable mentions) that are linked to 11,356 variables of which 1,165 are unique. The shared task is divided into two sub-tasks: a) variable detection and b) variable disambiguation. The former deals with identifying sentences that contain variable mentions, while the latter focuses on linking the correct variables mentioned in a sentence. Results show that implicit variables, which require contextual knowledge, are significantly more difficult to identify. Furthermore, we find that both tasks can be conducted in a zero-shot setting using pre-trained language models.

[4] https://vadis-project.github.io/sv-ident-sdp2022/

The SV-Ident overview paper is available at (Tsereteli et al., 2022).

4.4 Scholarly Knowledge Graph Generation (SKGG)

Organizers: Petr Knoth, David Pride, Ronin Wu and Drahomira Herrmannova

With the demise of the widely used Microsoft Academic Graph (MAG) (Wang et al., 2020; Herrmannova and Knoth, 2016) at the end of 2021, the scholarly document processing community faces a pressing need to replace MAG with an open source community supported service. A number of challenging data processing tasks are needed to create a comprehensive scholarly graph, i.e., a graph of entities including research papers, authors, research organisations, and research themes. This shared task aimed to evaluate three key sub-tasks of scholarly graph generation: 1) *document deduplication*, identifying and linking different versions of the same paper, 2) *extracting research themes*, and 3) *affiliation mining*, linking papers to the organisations that produced them. Unfortunately, participants only submitted results in the first subtask, using a new 50k large dataset of 36 research themes compiled based on the UK Research Excellence Framework exercise and enriched using the CORE (Knoth and Zdrahal, 2012) and the Semantic Scholar (Ammar et al., 2018) APIs. The task has created a new performance benchmark comparing traditional and state-of-the-art models under the same experimental conditions. The highest performance was achieved by a transformer-based classifier model based on BERT with the use of argumentative zoning. The SKGG overview paper is available at Óscar E. Mendoza et al. (2022).

4.5 Multi Perspective Scientific Document Summarization (MuP 2022)

Organizers: Arman Cohan, Guy Feigenblat, Tirthankar Ghosal and Michal Shmueli-Scheuer

MuP 2022 shared task is 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. Systems reports submitted by 5 teams are included in the workshop proceedings along with an overview paper summarizing results and insights.

While we observe encouraging results from the participating teams, we conclude that there is still significant room left for improving summarization leveraging multiple references. The MuP overview paper is available at Cohan et al. (2022).

5 Workshop Overview and Outlook

The organizers were gratified by both the size and breadth of the response to the third edition of SDP. The subjects of accepted papers ranged from end uses of the scholarly literature (such as search, document expansion, or writing support) to challenges associated with automated understanding (such as metadata extraction and disambiguation or argument mining), to adaptations of recent successes in the broader field of NLP. 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 COLING, it is perhaps not surprising that the majority of our submissions emphasized NLP. As we consider future iterations of the workshop, we are discussing ways to increase its subject diversity. With SDP 2022 we have begun to present 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.

6 Conclusion

The scholarly literature has long served as a rich source of interesting and challenging problems for computer science, and there is substantial prior work in information retrieval, scientometrics, data mining, and computational linguistics, but many important challenges remain. In many

respects, our efforts to faithfully capture the semantics of scholarly communication through automated means are still in their infancy. At the same time, recent events regarding misinterpretation of scholarly information accentuate the importance of better approaches to the automated processing of scholarly literature.

By drawing attention to these problems and offering a forum for interested scientists from a range of disciplines to collaborate, we hope that this and future instances of SDP encourage the application of recent advances in relevant fields to this problem area, identify new use cases or improve our understanding of existing ones, and ultimately foster solutions that improve the practice of scholarship and serve society.

7 Program Committee

1. Akiko Aizawa, National Institute of Informatics, Japan
2. Hamed Alhoori, Northern Illinois University, USA
3. Iana Atanassova, Université de Bourgogne Franche-Comté, France
4. Premjith B, Amrita Vishwa Vidyapeetham, Coimbatore, India
5. Arie Cattan, Bar Ilan University, Israel
6. Yimeng Dai, University of Melbourne, Australia
7. Sourish Dasgupta, Dhirubhai Ambani Institute of Information and Communication Technology, India
8. Jay DeYoung, Northeastern University, USA
9. Alexander Fabbri, Salesforce, USA
10. Zheng Gao, Amazon Alexa AI, USA
11. John Giorgi, University of Toronto, Canada
12. Paul Groth, University of Amsterdam, Netherlands
13. Daisuke Ikeda, Kyushu University, Japan
14. Roman Kern, Graz University of Technology, Austria
15. Valia Kordoni, Humboldt University Berlin, Germany
16. Xiangci Li, University of Texas at Dallas, USA
17. Yoshitomo Matsubara, University of California, Irvine, USA
18. Aakanksha Naik, Carnegie Mellon University, USA
19. David Pride, The Open University, UK
20. Terry Ruas, University of Wuppertal, Germany
21. Angelo Antonio Salatino, The Open University, UK
22. Zejiang Shen, Massachusetts Institute of Technology, USA
23. Mayank Singh, Indian Institute of Technology Gandhinagar, India
24. Neil Smalheiser, University of Illinois at Chicago, USA
25. Markus Stocker, German National Library of Science and Technology, Germany
26. Wojtek Sylwestrzak, University of Warsaw, Poland
27. Rajeev Verma, Indian Institute of Technology Patna, India
28. Boris Veytsman, Chan Zuckerberg Initiative, USA
29. David Wadden, University of Washington, USA
30. Byron Wallace, Northeastern University, USA
31. Xuan Wang, University of Illinois at Urbana-Champaign, USA
32. Jian Wu, Old Dominion University, USA
33. Wuhe Zou, NetEase AI, China

Acknowledgements

We organizers wish to thank all those who contributed to this workshop series: The researchers who contributed papers, the many reviewers who generously offered their time and expertise, and the participants of the workshop.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

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.

Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2021. Tortured phrases: A dubious writing style emerging in science. evidence of critical issues affecting established journals. *arXiv preprint arXiv:2107.06751*.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the First Workshop

on Scholarly Document Processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

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

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS2: Multi-document summarization of medical studies. In *EMNLP*.

Óscar E. Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, and Allan Hanbury. 2022. Benchmark for research theme classification of scholarly documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Drahomira Herrmannova and Petr Knoth. 2016. An analysis of the microsoft academic graph. *D-Lib Magazine*, 22(9/10).

Yury Kashnitsky, Drahomira Herrmannova, Anita de Waard, Georgios Tsatsaronis, Catriona Fennell, and Cyril Labbé. 2022. Overview of the DAGPap22 shared task on detecting automatically generated scientific papers. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Petr Knoth, Christopher Stahl, Bikash Gyawali, David Pride, Suchetha N. Kunnath, and Drahomira Herrmannova, editors. 2020. *Proceedings of the 8th International Workshop on Mining Scientific Publications*. Association for Computational Linguistics, Wuhan, China.

Petr Knoth and Zdenek Zdrahal. 2012. Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12).

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.

Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher. 2007. How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal medicine*, 147(4):224–233.

Tornike Tsereteli, Yavuz Selim Kartal, Simone Paolo Ponzetto, Andrea Zielinski, Kai Eckert, and Philipp Mayr. 2022. Overview of the SV-Ident 2022 Shared Task on Survey Variable Identification in Social Science Publications. In *Proceedings of the Third Workshop on Scholarly Document Processing*. Association for Computational Linguistics.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2020. Generating (factual?) narrative summaries of RCTs: Experiments with neural multi-document summarization. In *AMIA Annual Symposium*.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.