Overview of the Fourth Workshop on Scholarly Document Processing

Tirthankar Ghosal[a] **Amanpreet Singh**[b] **Anita de Waard**[c] **Philipp Mayr**[d]
Aakanksha Naik[b] **Orion Weller**[e] **Yoonjoo Lee**[f] **Shannon Shen**[g] **Yanxia Qin**[h]

Abstract

The workshop on Scholarly Document Processing (SDP) started in 2020 to accelerate research, inform policy and educate the public on natural language processing for scientific text. The fourth iteration of the workshop, SDP24 was held at the 62nd Annual Meeting of the Association for Computational Linguistics (ACL24) as a hybrid event. The SDP workshop saw a great increase in interest, with 57 submissions, of which 28 were accepted. The program consisted of a research track, four invited talks and two shared tasks: 1) DAGPap24: Detecting automatically generated scientific papers and 2) Context24: Multimodal Evidence and Grounding Context Identification for Scientific Claims. The program was geared towards NLP, information extraction, information retrieval, and data mining for scholarly documents, with an emphasis on identifying and providing solutions to open challenges.

1 Workshop description

Scholarly literature is the chief means by which scientists and academics document and communicate their results and is therefore critical to the advancement of knowledge and improvement of human well-being.

At the same time, this literature poses challenges to NLP uncommon in other genres, such as specialized language and high background knowledge requirements, long documents and strong structural conventions, multi-modal presentation, citation relationships among documents, an emphasis on rational argumentation, and the frequent availability of detailed metadata. These challenges necessitate the development of NLP methods and resources optimized for this domain. The Scholarly Document Processing (SDP) workshop provides a venue for discussing these challenges, bringing together stakeholders from different communities including computational linguistics, machine learning, text mining, information retrieval, digital libraries, scientometrics and others, to develop methods, tasks, and resources in support of these goals.

In addition to the shared tasks (see Section 4), SDP invited research submissions on a diversity of topics ranging from technical challenges posed by scholarly texts to novel applications facilitating scholarly work.

The first **Scholarly Document Processing** (SDP) workshop was co-located online with the EMNLP 2020 conference (Chandrasekaran et al., 2020), and provided a dedicated venue for those working on SDP to submit and discuss their research. Following this succes and the clear appetite for venues to foster discussions around scholarly NLP, SDP 2021 co-located with NAACL (Beltagy et al., 2021), and SDP 2022 with COLING (Cohan et al., 2022) 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.

2 Program

The SDP 2024 workshop consisted of four keynote talks, a research track and a shared task track. SDP 2024 received 57 submissions, of which 28 were accepted (49% acceptance rate).

[a] Oak Ridge National Laboratory, USA
[b] Allen Institute for AI, USA
[c] Elsevier, Netherlands
[d] GESIS -- Leibniz Institute for the Social Sciences, Germany
[e] Johns Hopkins University, USA
[f] Korea Advanced Institute of Science & Technology, South Korea
[g] Massachussets Institute of Technology, USA
[h] National University of Singapore, Singapore

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 1–6
August 16, 2024 ©2024 Association for Computational Linguistics

Since the workshop will be hybrid, there will be both in-person and virtual presentations at the conference venue and online. Topics of the presentations run the gamut, and include representation learning of scientific documents (including tables and images), citation text as well as paper text generation, peer review processing, scientific document summarization, AI-generated text detection, scientific intuition and reasoning, scientific papers to proposals, entity recognition and author disambiguation, scientific claim verification, and many more. As expected, we see a sharp increase in papers that employ large language models for downstream SDP tasks. The full program with links to papers, videos and posters is available at https://sdproc.org/2024/program.html.

3 Keynotes

This year we had four keynote lectures from researchers who are prominent in Natural Language Processing for scholarly documents:

- **Doug Downey**, Associate Professor at Northwestern University and Research Director at Allen Institute for AI, USA.
- **Iryna Gurevych**, Professor at Technical University Darmstadt and head of the UKP Lab, Germany.
- **Anna Rogers**, Assistant Professor, University of Copenhagen, Denmark.
- **Heng Ji**, Professor, University of Illinois at Urbana-Champaign, USA.

Speaker Iryna Gurevych

Title “How to InterText? Elevating NLP to the cross-document level”

Abstract While modern language models do a great job at finding documents, extracting information from them and generating naturally sounding language, the progress in helping humans read, connect, and make sense of interrelated long texts has been very much limited. Funded by the European Research Council, the InterText project brings natural language processing (NLP) forward by developing a general framework for modeling and analyzing fine-grained relationships between texts – intertextual relationships. This crucial milestone for AI would allow tracing the origin and evolution of texts and ideas and enable a new generation of AI applications for text work and critical reading. Using the scientific domain as a prototypical model of collaborative knowledge construction anchored in text, this talk will provide an overview of UKP Lab’s past and ongoing research demonstrating our intertextual approach to NLP in the scientific domain. Specifically, we will highlight two lines of our work. The first one is related to task design, practical applications and intricacies of data collection in the peer-review domain. The second one is about scientific text generation targeting (i) citation text and (ii) attitude and theme-guided rebuttals. To conclude, we will briefly describe our ongoing efforts towardsfine-grained linking of multiple documents, temporal analysis of scientific datasets and research novelty modeling.

Speaker Heng Ji

Title “AI Plays Medicinal Chemist”

Abstract There exist approximately 166 billion small molecules, with 970 million deemed drug-like. Despite this vast pool, only 89 tyrosine kinase inhibitors are currently approved across global healthcare systems. This scarcity underscores the urgent need for innovative approaches, calling upon the NLP community to contribute significantly to medicine. However, the challenges are manifold. Existing large language models (LLMs) alone are insufficient due to their tendency to generate erroneous claims confidently (hallucinate). Moreover, traditional knowledge bases do not adequately address the issue; none of the 89 kinase inhibitors are documented in popular human-constructed databases. This gap persists because chemistry language diverges significantly from natural language, demanding specialized domain knowledge, multimodal information integration, and long context understanding. Using drug discovery as a case study, I will present our approaches to tackle these challenges and turn an AI agent into a Medicinal Chemist. I will share preliminary results from animal testing conducted on drug variants proposed by AI algorithms. Furthermore, I advocate for a paradigm shift towards ‘slow science’, emphasizing the integration of feedback loops from molecule synthesis and animal testing. This new paradigm aims to evaluate AI techniques in scientific contexts, moving beyond chasing precision/recall scores at leaderboards which are prevalent in the current computer science community.

Speaker Doug Downey

Title "Chasing high-precision NLP at discount prices: Lessons for accelerating science"

Abstract Natural language processing (NLP) has made major strides in recent years, due to the increasing capabilities of large language models. However, using NLP to power real applications is still challenging: the best models are expensive to apply at scale and are still prone to errors. I'll describe recent lessons we've learned on the Semantic Scholar team as we've built and deployed applications using NLP aimed at accelerating science, including PDF content extraction, automatically-constructed topic pages for science, and complex question answering. While recent NLP breakthroughs do enable exciting new experiences, fully delivering on the potential of this technology will require solving multiple open research problems.

Speaker Anna Rogers

Title "Large language models as research assistants: workflows and challenges"

Abstract Research practices in our and other fields are being actively reshaped by the new tools based on large language models. For every step in the traditional research pipeline, from experimentation to writing, commercial 'solutions' are already actively marketed. This talk will discuss to what extent the marketing is realistic, how the research practices seem to be changing, and how all this interacts with considerations of publication ethics and security.

4 Shared Task Track

SDP 2024 hosted two shared tasks. Both shared tasks had their own organizing committees consisting of several members of the SDP 2024 organizers and/or other collaborators. Detailed overview papers of the shared tasks are referred to and followed in the proceedings.

4.1 Detecting automatically generated scientific papers (DAGPap24)

Organizers: Savvas Chamezopoulos, Drahomira Herrmannova, Anita de Waard, Domenic Rosati, Yury Kashnitsky

A big problem with the ubiquity of Generative AI is that it has now become very easy to generate fake scientific papers. This can erode public trust in science and attack the foundations of science: are we standing on the shoulders of robots? The Detecting Automatically Generated Papers (DAGPap) competition aims to encourage the development of robust, reliable AI-generated scientific text detection systems, utilizing a diverse dataset and varied machine learning models in a number of scientific domains.

Building on the DagPap22 Competition Kashnitsky et al. (2022), this year's dataset consisted of 30,000 scientific articles sourced from ScienceDirect [1] that were processed to integrate various alteration methods within the human-written content. 28 teams participated in the task, with a total of 457 submissions. Out of these, 19 teams managed to beat the SciBERT baseline on the development set, and 12 teams managed to beat the SciBERT baseline on the test set. Four top teams exceeded an F1-score of 99%. For more details on the challenge, please see the DAGPap24 overview paper, elsewhere in this proceedings.

4.2 Multimodal Evidence and Grounding Context Identification for Scientific Claims (Context24)

Organizers: Joel Chan, Matt Akamatsu, and Aakanksha Naik

Interpreting scientific claims in the context of empirical findings is a valuable practice, yet extremely time-consuming for researchers. Such interpretation requires identifying key results (e.g., figures, tables, etc.) that provide supporting evidence from research papers, and contextualizing these results with associated methodological details (e.g., measures, sample, etc.). In this shared task, we released datasets to encourage the development of models for automatic identification of key results and additional grounding context, given a scientific claim. Context24 consisted of two tracks: (i) evidence identification, and (ii) grounding context identification. For track 1, we released a training dataset of 474 claims with key figure/table annotations, and manually extracted figures, tables and captions for all papers to avoid PDF parsing issues. For track 2, given the challenging annotation process, we released a limited training set of 42 claims to encourage exploration of zero-shot or few-shot methods. We also released full-texts for all associated papers

[1] https://www.sciencedirect.com/

parsed using PaperMage (Lo et al., 2023), and full-texts for ∼17,000 additional related papers (i.e., papers cited by/citing the original set) to encourage investigation of data augmentation and weak supervision-based approaches. Finally we released test sets consisting of 111 and 109 instances for tracks 1 and 2 respectively. For track 1, we received submissions from seven teams, while only two participated in track 2 possibly due to the low-resource nature of the task. We observed modest to strong performance improvements in both tracks as measured by automated evaluation metrics indicating that models find this task tractable but still have room to improve further. System reports from top three participating teams as well as an overview paper summarizing future directions (Chan et al., 2024) are included in the workshop proceedings.

5 Workshop Overview and Outlook

The organizers were gratified by both the size and breadth of the response to the fourth edition of SDP. The subjects of accepted papers ranged from end uses of the scholarly literature to challenges associated with automated understanding to adaptations of recent successes of LLMs 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 ACL, 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 2024 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 meta-linguistic 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. Abdelhalim Hafedh DAHOU, Universität Ulm
2. Abhinav Ramesh Kashyap, National University of Singapore
3. Akiko Aizawa, NII, Tokyo Institute of Technology
4. Alexander Fabbri, SalesForce.com
5. Alexander Shirnin, Higher School of Economics
6. Allan Hanbury, Complexity Science Hub and Technische Universität Wien
7. Allen G Roush, Oracle
8. Angelo Salatino, KMI - Open University
9. Anurag Acharya, Pacific Northwest National Laboratory
10. Arman Cohan, Yale University and Allen Institute for Artificial Intelligence
11. Ashok Urlana, Tata Consultancy Services Limited, India
12. Autumn Toney, Georgetown University
13. Biswadip Mandal, University of Texas at Dallas
14. Boris Veytsman, Chan Zuckerberg Initiative and George Mason University
15. Brian Douglas Zimmerman, University of Waterloo
16. Buse Sibel Korkmaz, Imperial College London
17. Daisuke Ikeda, Kyushu University and Kyushu

University
18. Dana Moukheiber, Massachusetts Institute of Technology
19. Daniel Acuna, Computer Science Department, University of Colorado at Boulder
20. Danilo Dessi, GESIS
21. Dayne Freitag, SRI International
22. Debarshi Kumar Sanyal, Indian Association for the Cultivation of Science
23. Doug Downey, Allen Institute for Artificial Intelligence and Northwestern University
24. Drahomira Herrmannova, Elsevier
25. Faiza BELBACHIR, Institut polytechnique des sciences avancees
26. German Gritsai, Université Grenoble Alpes
27. Halil Kilicoglu, University of Illinois at Urbana-Champaign
28. Hamed Alhoori, Northern Illinois University
29. Hiroki Teranishi, Nara Institute of Science and Technology, Japan and RIKEN
30. Hosein Azarbonyad, Elsevier
31. Ibrahim Al Azher, Northern Illinois University
32. Ioana Buhnila, Université de Lorraine
33. James Dunham, Georgetown University
34. Jan Philip Wahle, University of Göttingen, Germany
35. Jay DeYoung, Northeastern University
36. John Michael Giorgi, Toronto University
37. Julio C. Rangel, RIKEN
38. Jun Zhuang, Boise State University and Indiana University Purdue University Indianapolis
39. Kazuya Nishimura, Kyushu University
40. Kiran Sharma, BML Munjal University
41. Kyle Lo, Allen Institute for Artificial Intelligence
42. Lars Benedikt Kaesberg,
43. Lucy Lu Wang, University of Washington and Allen Institute for Artificial Intelligence
44. Markus Stocker, TIB Leibniz Information Centre for Science and Technology
45. Miftahul Jannat Mokarrama, Northern Illinois University
46. Naaman Tan, National University of Singapore
47. Neil R. Smalheiser, University of Illinois at Chicago
48. Nianlong Gu, University of Zurich
49. Nicolau Duran-Silva, Universitat Pompeu Fabra
50. Nikita Andreev, Yandex
51. Nina Smirnova, GESIS – Leibniz Institute for the Social Sciences
52. Peter Zhang, University of California, Berkeley
53. Petr Knoth, Open University
54. Pierre Senellart, Ecole Normale Supérieure
55. Roman Kern, Know Center GmbH and Technische Universität Graz
56. Sameera Horawalavithana, Pacific Northwest National Laboratory
57. Sandeep Kumar, Indian Institute of Technology, Patna
58. Sebastian Schellhammer, GESIS Leibniz Institute for the Social Sciences
59. Sharmila upadhyaya, Gesis leibniz Institute
60. Shiyuan Zhang, University of Illinois at Urbana-Champaign
61. Shufan Ming, University of Illinois at Urbana-Champaign
62. Soham Chitnis, Birla Institute of Technology and Science Pilani
63. Sotaro Takeshita, Universität Mannheim
64. Sourish Dasgupta, Dhirubhai Ambani Institute of Information & Communication Technology
65. Tamjid Azad, Northern Illinois University
66. Taro Watanabe, Nara Institute of Science and Technology, Japan
67. Terry Ruas, Georg-August Universität Göttingen
68. Tianyong Hao, South China Normal University
69. Tim Schopf, Technische Universität München
70. Tohida Rehman, Jadavpur University
71. Tom Hope, Allen Institute for Artificial Intelligence and Hebrew University, Hebrew University of Jerusalem
72. Tornike Tsereteli, Universität Mannheim
73. Tosho Hirasawa, Omron Sinic X
74. Vladislav Mikhailov, University of Oslo
75. Wojciech Kusa, Allegro
76. Wojtek Sylwestrzak, University of Warsaw
77. Wolfgang Otto, GESIS
78. Xiangci Li, Google
79. Xiaoliang Jiang, University of Illinois at Urbana-Champaign
80. Xinyuan Lu, national university of singaore, National University of Singapore
81. Yagmur Ozturk, Université Grenoble Alpes
82. Yavuz Selim Kartal, Gesis
83. Yixi Ding, National University of Singapore
84. Yoshitomo Matsubara, Spiffy AI
85. Yupeng Cao, Stevens Institute of Technology
86. Yury Kashnitsky, Google
87. Zoran Medić, UniZg-FER, University of Zagreb

88. Wuhe Zou, Netease Group

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.

Philipp Mayr received funding via the OUTCITE project (Backes et al., 2024) at GESIS Cologne and University Stuttgart (Funder: Deutsche Forschungsgemeinschaft; Grant: MA 3964/8-2).

References

Tobias Backes, Anastasiia Iurshina, Muhammad Ahsan Shahid, and Philipp Mayr. 2024. Comparing Free Reference Extraction Pipelines. *International Journal on Digital Libraries*.

Iz Beltagy, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Keith Hall, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Robert Patton, Michal Shmueli-Scheuer, Anita de Waard, Kuansan Wang, and Lucy Lu Wang. 2021. Overview of the second workshop on scholarly document processing. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 159–165, Online. Association for Computational Linguistics.

Chu Sern Joel Chan, Aakanksha Naik, Matt Akamatsu, Hanna Bekele, Erin Bransom, Ian Campbell, and Jenna Sparks. 2024. Overview of the context24 shared task on contextualizing scientific claims. In *The 4th Workshop on Scholarly Document Processing @ ACL 2024*.

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, Dayne Freitag, Tirthankar Ghosal, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Michal Shmueli-Scheuer, Anita de Waard, and Lucy Lu Wang. 2022. Overview of the third workshop on scholarly document processing. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 1–6, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. 2023. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore. Association for Computational Linguistics.