Overview of the Second Workshop on Scholarly Document Processing

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

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 2nd Workshop on Scholarly Document Processing (SDP) at NAACL 2021 as a virtual event (https://sdproc.org/2021/). The SDP workshop consisted of a research track, three invited talks and three Shared Tasks (LongSumm 2021, SCIVER and 3C). 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.

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 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.

We believe that ACL events are the most appropriate venue for the SDP workshop for three 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 furthering many tasks and applications in scholarly document processing. Second, the clear practical importance

[a]Allen Institute for AI, USA
[b]IBM Research AI, Haifa Research Lab, Israel
[c]SRI International, USA
[d]ÚFAL, Charles University, Czech Republic
[e]Google AI, USA
[f]Oak Ridge National Laboratory, USA
[g]The Open University, UK
[h]GESIS – Leibniz Institute for the Social Sciences, Germany
[i]Elsevier, USA
[j]Microsoft Research, USA

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

Proceedings of the Second Workshop on Scholarly Document Processing, pages 159–165

of the scholarly literature makes it an attractive testbed and source of distinctive challenges for researchers focused more generally on computational linguistics. By co-locating with ACL, we aim to expand the SDP community by drawing the attention of computational linguists and NLP researchers in search of important, practical problem areas. And third, we seek to bring together researchers and practitioners from various backgrounds focusing on different aspects of scholarly document processing. We believe that the interdisciplinary nature of the ACL venues greatly assists in encouraging submissions from a diverse set of fields.

Topics 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.

We specifically invited research on important and under-served practical needs, such as:

1. Identifying/mitigating scientific disinformation and its effects on public policy and behavior,
2. Reducing information overload through summarization and aggregation of information within and across documents, and
3. Improving access to scientific papers through multilingual scholarly document processing.

Program The SDP 2021 workshop consisted of three Keynote talks, a Research Track and a Shared Task Track with three separate tasks: LongSumm, SCIVER, and 3C. The full program with links to papers, videos and posters is available at https://sdproc.org/2021/program.html.

2 Keynotes

(1) **Yoav Goldberg**, Professor, Bar Ilan University gave the first keynote titled: "Empowering Experts with Extractive Search". A recording of his talk can be found on YouTube.

Talk abstract: "Digitization and search has revolutionized information access. Yet, current search systems are all geared towards a specific kind of information need. The majority of systems are precision oriented, getting you the most relevant documents on a given topic. Some expert-oriented system are recall focused, and aim to find all documents on a given topic. Some systems provide snippets rather than full documents, and recent advances in QA allow to highlight the spans where the answer may be. In all these cases, the user needs to look at all the returned answers and process them themselves. This works very well if the answer you are looking for is written in a single, or a few, documents. But not all information needs are like that. We present a different kind of search system, which is geared toward answering information needs whose answers are based on aggregation of pieces of information over a large corpus. The key component is allowing users to define query elements that act as variables, or "captures", which are then extracted from each matching result, and presented in aggregation. This allows us to formulate queries to answer questions such as "what are various ways of referring to leprosy", "what are common reported incubation period for COVID-19", "what are the kinds of treatments considered in the literature for Alzheimer's disease", "what is being coated by fibronectin" and so on. We demonstrate SPIKE, a publicly available prototype of such an extractive search system, and discuss its current capabilities, and also limitations and future directions."

(2) **Isabelle Augenstein**, Associate Professor, University of Copenhagen gave the second keynote titled: "Determining the Credibility of Science Communication". An outline of her talk can be found in the extended abstract (Augenstein, 2021).

Talk abstract: "Most work on scholarly document processing assumes that the information processed is trustworthy and factually correct. However, this is not always the case. There are two core challenges, which should be addressed: 1) ensuring that scientific publications are credible

– e.g. that claims are not made without supporting evidence, and that all relevant supporting evidence is provided; and 2) that scientific findings are not misrepresented, distorted or outright misreported when communicated by journalists or the general public. I will present some first steps towards addressing these problems and outline remaining challenges."

(3) **Hannaneh Hajishirzi**, Assistant Professor, University of Washington gave the third keynote titled: "Knowledge Acquisition from Unstructured Scientific Text". A recording of her talk can be found on YouTube.

Talk abstract: "Enormous amounts of ever-changing knowledge are available online in diverse emergent textual styles (e.g., news vs. science text). Recent advances in deep learning algorithms, large-scale datasets, and industry-scale computational resources are spurring progress in many Natural Language Processing (NLP) tasks. Nevertheless, current models lack the ability to understand emergent domains such as scientific articles related to COVID-19 when training data are scarce. This talk presents some of recent efforts in our lab to address the problem of textual comprehension and reasoning about scientific articles. First, I discuss our multi-task learning approach for identifying and classifying entities and their relations in scientific articles. I further show how we can extend this approach to extract mechanism relations from COVID-19 articles to construct a scientific knowledge graph, which supports advanced search for medical doctors. Second, I introduce scientific claim verification, a new task to select abstracts from the research literature containing evidence that supports or refutes a given scientific claim, and to identify rationales justifying each decision. I finally show that our claim verification system is able to identify plausible evidence for 70% claims relevant to COVID-19 on the CORD-19 corpus."

3 Research Track

In total, we received 26 submissions for the research track. We accepted 11 papers for presentation (5 as long papers and 6 as short papers). One accepted paper was withdrawn by the authors after notification. We rejected 15 research paper submissions.

The accepted papers are:

Long papers:

- Javier Corvi, Carla Fuenteslópez, José Fernández, Josep Gelpi, Maria-Pau Ginebra, Salvador Capella-Guitierrez and Osnat Hakimi: *The Biomaterials Annotator: a system for ontology-based concept annotation of biomaterials text*.
- Ibrahim Burak Ozyurt, Joseph Menke, Anita Bandrowski and Maryann Martone: *Detecting Anatomical and Functional Connectivity Relations in Biomedical Literature via Language Representation Models*.
- Soyeong Jeong, Jinheon Baek, ChaeHun Park and Jong Park: *Unsupervised Document Expansion for Information Retrieval with Stochastic Text Generation*.
- Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami and Hirotoshi Taira: *Task Definition and Integration For Scientific-Document Writing Support*.

Short papers:

- Athar Sefid, Prasenjit Mitra, Jian Wu and C Lee Giles: *Extractive Research Slide Generation Using Windowed Labeling Ranking*.
- Johan Krause, Igor Shapiro, Tarek Saier and Michael Färber: *Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic*.
- Lee Kezar and Jay Pujara: *Finding Pragmatic Differences Between Disciplines*.
- Yash Gupta, Pawan Sasanka Ammanamanchi, Shikha Bordia, Arjun Manoharan, Deepak Mittal, Ramakanth Pasunuru, Manish Shrivastava, Maneesh Singh, Mohit Bansal and Preethi Jyothi: *The Effect of Pretraining on Extractive Summarization for Scientific Documents*.
- Chrysovalantis Giorgos Kontoulis, Eirini Papagiannopoulou and Grigorios Tsoumakas: *Keyphrase Extraction from Scientific Articles via Extractive Summarization*.
- Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard and Dayne Freitag: *Argument Mining for Scholarly Document Processing: Taking Stock and Looking Ahead*.

4 Shared Task Track

SDP 2021 hosted three shared tasks – LongSumm, SCIVER, and 3C. Each shared task had its own organizing committee consisting of several members of the SDP 2021 organizers and/or other collaborators. Shared task presentations were held online in parallel sessions to the main SDP workshop.

4.1 LongSumm

The 2nd Shared Task on Generating Long Summaries for Scientific Documents (*LongSumm*). The task is fundamentally different than generating short summaries that mostly aim at teasing the reader. The LongSumm task strives to learn how to cover the salient information conveyed in a given scientific document, taking into account the characteristics and the structure of the text. The motivation for LongSumm was first demonstrated by the IBM Science Summarizer system, (Erera et al., 2019) that retrieves and creates long summaries of scientific documents[4]. While Erera et al. (2019) studied some use-cases and proposed a summarization approach with some human evaluation, the authors stressed the need of a large dataset that will unleash the research in this domain. LongSumm aims at filling this gap by providing large dataset of long summaries which are based on blogs written by Machine Learning and NLP experts.

4.1.1 Corpus and Evaluation Metrics

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. The evaluation was conducted using the ROUGE measure (Lin, 2004) and executed on a public leaderboard [5] forked from EvalAI [6]. In addition, a subset of randomly selected summaries, of the top ranked systems, was evaluated by experts. The dataset as well as the results are available on the LongSumm Github Page.

[4] https://ibm.biz/sciencesum
[5] https://aieval.draco.res.ibm.com/challenge/39/
[6] https://eval.ai/

4.1.2 Systems Overview

Six systems participated in the task this year, with a total of around 200 submissions. Three teams submitted peer-reviewed technical reports, that are published as part of the workshop proceedings. Similar to last year, the more advanced methods employed deep learning techniques to generate abstractive and extractive summaries. The more basic methods employed graph centrality or utilized the distribution of words in documents to select salient sentences for extractive summarization. The winning team this year is the N&E team from NetEase. The team has obtained the highest average of ROUGE F-1 scores and the highest rank on the human evaluation.

4.1.3 Discussion

Scientific documents can be characterized as long, structured, utilizing technical language (i.e., formulas, tables, definitions, etc.). Analyzing the summaries and reports of the participated systems shows that most of them considered the structure of the document while generating summaries, by utilizing sections and document discourse. Eliminating some sections could help in focusing the summary (e.g., abstract). However, it should be done carefully, and it is important to make sure that important sections are not ignored. Scientific documents often contain special entities including mathematical definitions, formulas, tables, and the text surrounding them. The entities are usually not textual, however, they have an important aspect in articulating and explaining important aspects from the document. Thus in the future, we might want to extend the task definition and evaluation to support such entities. Finally, readability should play an important role in algorithmic design. Due to the nature of scientific documents and LongSumm length requirement, we believe this is even more challenging compared to traditional summarization tasks. This should have gotten more attention by the participating systems.

4.2 SCIVER

Due to the rapid growth in scientific literature and the proliferation of mis- and dis-information about scientific facts online, there is a need for AI systems that can support automated verification of scientific claims and fact-checking using evidence found in the research literature.

With this goal in mind, the SCIVER shared task provided a platform for facilitating community ex-

ploration of different approaches in developing AI systems that can take a scientific claim as input, verify it as Supported or Refuted by research papers *and* provide evidenciary sentences, or "rationales," for the predicted labels. The shared task used the SciFact (Wadden et al., 2020) dataset of 1.4K expert-curated claims matched to relevant biomedical paper abstracts with Support/Refute labels and expert-annotated evidentiary sentences. Using this data, 11 teams made a total of 14 submissions to the shared task. The best systems improved upon the previous state-of-the-art baseline by +23 F1, demonstrating impressive advancement on this important task.

A complete overview of the task is available at Wadden and Lo (2021) along with summarized insights and findings on the most promising modeling approaches from the shared task, such as neural refinement of bag-of-words retrieval candidates, training with negative sampling, incorporating full-document context in sentence-level prediction, and adoption of large pre-trained language models. The shared task leaderboard will remain open for submissions at https://leaderboard.allenai.org/scifact.

4.3 Citation Context Classification (3C)

To address the problem of all citations being treated equally by research metrics and assist automated analysis of scientific literature, we ran a shared task focused on the development of models capable of identifying citation intent – the second Citation Context Classification (3C) shared task. The participating teams were provided with 3,000 citation sentences from the ACT dataset (Pride and Knoth, 2020) annotated with two labels: a purpose label ("background", "compares/contrasts", "extension", "future", "motivation", "uses"; subtask A) and an influence label ("incidental", "influential"; subtask B). This year, the participating teams were also provided with the full text content of all documents in both the training and the test set which was extracted from CORE (Knoth and Zdrahal, 2012).

This edition of the 3C Shared task witnessed the active participation of a larger number of teams compared to the first edition of the shared task (Kunnath et al., 2020) – the total of 27 teams participated in this year's 3C shared task, 14 of those teams participated in both subtasks. The overall scores obtained for both tasks also showed considerable improvement when compared to the previous year's results – the highest macro F1-scores reported for subtask A and B were 0.2697[7] and 0.6003[8], respectively, compared to 0.2056 and 0.5557 reported in 3C 2020 (Kunnath et al., 2020). A detailed description of the shared task and the results is provided in (Kunnath et al., 2021) and the leaderboard for both subtasks can be accessed via Kaggle[7,8].

5 Workshop Overview and Outlook

The organizers were gratified by both the size and breadth of the response to the second 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 co-location with NAACL, 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 2021 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

[7] https://www.kaggle.com/c/3c-shared-task-purpose-v2/leaderboard

[8] https://www.kaggle.com/c/3c-shared-task-influence-v2/leaderboard

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.

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.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

Isabelle Augenstein. 2021. Determining the credibility of science communication. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

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.

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.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. 2019. A summarization system for scientific documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*.

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):1–13.

Suchetha N Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. 2020. Overview of the 2020 WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 75–83. Association for Computational Linguistics.

Suchetha N Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. 2021. Overview of the 2021 SDP 3C citation context classification shared task. In *Proceedings of the Second Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

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.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Virtual - China.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *EMNLP*.

David Wadden and Kyle Lo. 2021. Overview and Insights from the SCIVER Shared Task on Scientific Claim Verification. In *Proceedings of the Second Workshop on Scholarly Document Processing*. Association for Computational Linguistics.