SDP 2024

The Fourth Workshop on Scholarly Document Processing

Proceedings of the Workshop

August 16, 2024

The SDP organizers gratefully acknowledge the support from the following sponsors.

Best Paper Award

Best Paper Award (DAGPap shared task)

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-151-3

Message from the SDP 2024 Organizing Committee

Welcome to the Fourth Workshop on Scholarly Document Processing (SDP) at ACL 2024.

As the body of scholarly literature grows, automated methods in NLP, text mining, information retrieval, document understanding etc. are needed to address issues of information overload, disinformation, reproducibility, and more. Though progress has been made, there are significant unique challenges to processing scholarly text that require dedicated attention. The goal of the Scholarly Document Processing series of workshops is to provide a venue for addressing these challenges, as well as a platform for tasks and resources supporting the processing of scientific documents. Our long-term objective is to establish scholarly and scientific texts as an essential domain for NLP research, to supplement current efforts on web text and news articles.

This workshop builds on the success of prior workshops: the SDP workshop held at COLING in 2022, NAACL in 2021 and EMNLP in 2020, and the SciNLP workshop held at AKBC 2020 and AKBC 2021. As in previous years, we have sought 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.

Organizing Committee

Program Chairs

Tirthankar Ghosal, Oak Ridge National Laboratory, USA
Amanpreet Singh, Allen Institute for AI, USA
Anita de Waard, Elsevier, Netherlands
Philipp Mayr, GESIS - Leibniz Institute for the Social Sciences, Germany
Aakanksha Naik, Allen Institute for AI, USA
Orion Weller, Johns Hopkins University, USA
Yoonjoo Lee, KAIST, Korea
Shannon Shen, Massachussets Institute of Technology, USA
Yanxia Qin, National University of Singapore, Singapore

Steering Committee members

Arman Cohan, Yale University, USA
Dayne Freitag, SRI International, USA
Tom Hope, Hebrew University of Jerusalem, Israel
Petr Knoth, Open University, UK
Kyle Lo, Allen Institute for AI, USA
Lucy Lu Wang, University of Washington, USA

Program Committee

Reviewers

Anurag Acharya, Daniel Acuna, Akiko Aizawa, Hamed Alhoori, Nikita Andreev, Tamjid Azad, Hosein Azarbonyad, Ibrahim Al Azher

Faiza Belbachir, Ioana Buhnila

Yupeng Cao, Soham Chitnis, Arman Cohan

Abdelhalim Hafedh Dahou, Sourish Dasgupta, Sourish Dasgupta, Anita De Waard, Jay DeYoung, Danilo Dessi, Yixi Ding, Doug Downey, James Dunham, Nicolau Duran-Silva

Alexander Fabbri, Dayne Freitag

Tirthankar Ghosal, John Michael Giorgi, German Gritsai, Nianlong Gu

Allan Hanbury, Tianyong Hao, Drahomira Herrmannova, Tosho Hirasawa, Tom Hope, Sameera Horawalavithana

Daisuke Ikeda

Xiaoliang Jiang

Lars Benedikt Kaesberg, Yavuz Selim Kartal, Yury Kashnitsky, Roman Kern, Halil Kilicoglu, Petr Knoth, Buse Sibel Korkmaz, Sandeep Kumar, Wojciech Kusa

Xiangci Li, Kyle Lo, Xinyuan Lu

Biswadip Mandal, Yoshitomo Matsubara, Philipp Mayr, Zoran Medić, Vladislav Mikhailov, Shufan Ming, Miftahul Jannat Mokarrama, Dana Moukheiber

Aakanksha Naik, Kazuya Nishimura

Wolfgang Otto, Yagmur Ozturk

Yanxia Qin

Abhinav Ramesh Kashyap, Julio C. Rangel, Tohida Rehman, Allen G Roush, Terry Ruas

Angelo Salatino, Debarshi Kumar Sanyal, Sebastian Schellhammer, Tim Schopf, Pierre Senellart, Kiran Sharma, Zejiang Shen, Alexander Shirnin, Amanpreet Singh, Neil R. Smalheiser, Nina Smirnova, Markus Stocker, Wojtek Sylwestrzak

Sotaro Takeshita, Naaman Tan, Hiroki Teranishi, Autumn Toney, Tornike Tsereteli

Sharmila Upadhyaya, Ashok Urlana

Boris Veytsman

Jan Philip Wahle, Lucy Lu Wang, Taro Watanabe, Orion Weller

Peter Zhang, Shiyuan Zhang, Jun Zhuang, Brian Douglas Zimmerman, Wuhe Zou

Keynote Talk
How to InterText? Elevating NLP to the cross-document level

Iryna Gurevych
Technical University Darmstadt and head of the UKP Lab

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.

Bio: Iryna Gurevych is a German computer scientist. She is Professor at the Department of Computer Science of the Technical University of Darmstadt and Director of Ubiquitous Knowledge Processing Lab. She has a strong background in information extraction, semantic text processing, machine learning and innovative applications of NLP to social sciences and humanities.Iryna Gurevych has published over 300 publications in international conferences and journals and is member of programme and conference committees of more than 50 high-level conferences and workshops (ACL, EACL, NAACL, etc.). She is the holder of several awards, including the Lichtenberg-Professorship Career Award und the Emmy-Noether Career Award (both in 2007). In 2021 she received the first LOEWE-professorship of the LOEWE programme. She has been selected as a ACL Fellow 2020 for her outstanding work in natural language processing and machine learning and is the Vice-president-elect of the ACL since 2021.

Keynote Talk
AI Plays Medicinal Chemist

Heng Ji
University of Illinois at Urbana-Champaign

Abstract: There exist approximately 166 billion small molecules, with 970 million deemed druglike. 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.

Bio: Heng Ji is a professor at Computer Science Department of University of Illinois at Urbana-Champaign. She received her B.A. and M. A. in Computational Linguistics from Tsinghua University, and her M.S. and Ph.D. in Computer Science from New York University. Her research interests focus on Natural Language Processing, especially on Information Extraction and Knowledge Base Population. She is selected as Young Scientist and a member of the Global Future Council on the Future of Computing by the World Economic Forum in 2016 and 2017. The awards she received include AI's 10 to Watch Award by IEEE Intelligent Systems in 2013, NSF CAREER award in 2009, PACLIC2012 Best paper runner-up, 'Best of ICDM2013' paper award, Best of SDM2013 paper award, ACL2018 Best Demo paper nomination, Google Research Award in 2009 and 2014, IBM Watson Faculty Award in 2012 and 2014 and Bosch Research Award in 2014-2018. She has coordinated the NIST TAC Knowledge Base Population task since 2010. She is the associate editor for IEEE/ACM Transaction on Audio, Speech, and Language Processing. She has served as the Program Committee Co-Chair of NAACL-HLT2018, NLP-NABD2018, NLPCC2015, CSCKG2016 and CCL2019, and senior area chair for many conferences. She has led several multi-institute research efforts including DARPA DEFT Tinker Bell team of seven universities and DARPA KAIROS RESIN team of six universities. She is the task leader of the U.S. ARL projects on information fusion and knowledge networks construction between 2009-2019. She is invited by the Secretary of the Air Force and AFRL to join Air Force Data Analytics Expert Panel to inform the Air Force Strategy 2030.

Keynote Talk

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

Doug Downey
Northwestern University and Allen Institute for AI, USA

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.

Bio: Doug is a Research Director at AI2. He is currently on leave from Northwestern University, where he is an Associate Professor of Computer Science. His research focuses on information extraction, natural language processing, and machine learning. Outside of work, he enjoys spending time with family, exploring the outdoors, and watching movies.

Keynote Talk

Large language models as research assistants: workflows and challenges

Anna Rogers
University of Copenhagen

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.

Bio: Anna Rogers is an assistant professor in the Center for Social Data Science at the University of Copenhagen. She is currently also a visiting researcher with the RIKEN Center for Computational Science (Japan). Her main research area is Natural Language Processing, in particular model analysis and evaluation of natural language understanding systems.

Table of Contents

Program

Friday, August 16, 2024

08:30 - 17:30	*For the final SDP2024 Program Schedule, see the workshop website: https://sdproc.org/2024/.*