

COLING

**International Conference on
Computational Linguistics**

Proceedings of the Conference and Workshops

COLING

Volume 29 (2022), No. 9

Proceedings of the Workshop

Third Workshop on Scholarly Document Processing

October 12 - 17, 2022
Gyeongju, Republic of Korea

Copyright of each paper stays with the respective authors (or their employers).

ISSN 2951-2093

Message from the SDP 2022 Organizing Committee

Welcome to the Third Workshop on Scholarly Document Processing (SDP) at COLING 2022.

The SDP workshop has existed in other forms over the years, mainly in digital libraries or information sciences venues. In recent years, we have transitioned to organizing the SDP workshop at ACL events for several 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 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 events, we aimed 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 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

Arman Cohan, Allen Institute for Artificial Intelligence, USA

Guy Feigenblat, Piiano Privacy Solutions, Israel

Dayne Freitag, SRI International, San Diego, USA

Tirthankar Ghosal, Institute of Formal and Applied Linguistics, Charles University, Czech Republic

Drahomira Herrmannova, Elsevier, USA

Petr Knoth, Open University, UK

Kyle Lo, Allen Institute for Artificial Intelligence, USA

Philipp Mayr, GESIS – Leibniz Institute for the Social Sciences, Germany

Michal Shmueli-Scheuer, IBM Research AI, Haifa Research Lab, Israel

Anita de Waard, Elsevier, USA

Lucy Lu Wang, Allen Institute for Artificial Intelligence and University of Washington, USA

Table of Contents

<i>Overview of the Third Workshop on Scholarly Document Processing</i> 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	1
<i>Finding Scientific Topics in Continuously Growing Text Corpora</i> André Bittermann and Jonas Rieger	7
<i>Large-scale Evaluation of Transformer-based Article Encoders on the Task of Citation Recommendation</i> Zoran Medić and Jan Snajder	19
<i>Investigating the detection of Tortured Phrases in Scientific Literature</i> PUTHINEATH LAY, Martin Lentschat and Cyril Labbe	32
<i>Lightweight Contextual Logical Structure Recovery</i> Po-Wei Huang, Abhinav Ramesh Kashyap, Yanxia Qin, Yajing Yang and Min-Yen Kan	37
<i>Citation Context Classification: Critical vs Non-critical</i> Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé and François Portet	49
<i>Incorporating the Rhetoric of Scientific Language into Sentence Embeddings using Phrase-guided Distant Supervision and Metric Learning</i> Kaito Sugimoto and Akiko Aizawa	54
<i>Identifying Medical Paraphrases in Scientific versus Popularization Texts in French for Laypeople Understanding</i> Ioana Buhnila	69
<i>Multi-objective Representation Learning for Scientific Document Retrieval</i> Mathias Parisot and Jakub Zavrel	80
<i>Visualisation Methods for Diachronic Semantic Shift</i> Raef Kazi, Alessandra Amato, Shenghui Wang and Doina Bucur	89
<i>Unsupervised Partial Sentence Matching for Cited Text Identification</i> Kathryn Ricci, Haw-Shiuan Chang, Purujit Goyal and Andrew McCallum	95
<i>Multi-label Classification of Scientific Research Documents Across Domains and Languages</i> Autumn Toney and James Dunham	105
<i>Investigating Metric Diversity for Evaluating Long Document Summarisation</i> Cai Yang and Stephen Wan	115
<i>Exploiting Unary Relations with Stacked Learning for Relation Extraction</i> Yuan Zhuang, Ellen Riloff, Kiri L. Wagstaff, Raymond Francis, Matthew P. Golombek and Leslie K. Tamppari	126
<i>Mitigating Data Shift of Biomedical Research Articles for Information Retrieval and Semantic Indexing</i> Nima Ebadi, Anthony Rios and peyman najafirad	138
<i>A Japanese Masked Language Model for Academic Domain</i> Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai and Takashi Ninomiya	152

<i>Named Entity Inclusion in Abstractive Text Summarization</i>	
Sergey Berezin and Tatiana Batura	158
<i>Named Entity Recognition Based Automatic Generation of Research Highlights</i>	
Tohida Rehman, Debarshi Kumar Sanyal, Prasenjit Majumder and Samiran Chattopadhyay ...	163
<i>Citation Sentence Generation Leveraging the Content of Cited Papers</i>	
Akito Arita, Hiroaki Sugiyama, Kohji Dohsaka, Rikuto Tanaka and Hirotoishi Taira	170
<i>Overview of MSLR2022: A Shared Task on Multi-document Summarization for Literature Reviews</i>	
Lucy Lu Wang, Jay DeYoung and Byron Wallace	175
<i>LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation</i>	
Yulia Otmakhova, Tinh Hung Truong, Timothy Baldwin, Trevor Cohn, Karin Verspoor and Jey Han Lau	181
<i>Evaluating Pre-Trained Language Models on Multi-Document Summarization for Literature Reviews</i>	
Benjamin Yu	188
<i>Exploring the limits of a base BART for multi-document summarization in the medical domain</i>	
Ishmael Obonyo, Silvia Casola and Horacio Saggion	193
<i>Abstractive Approaches To Multidocument Summarization Of Medical Literature Reviews</i>	
Rahul Tangsali, Aditya Jagdish Vyawahare, Aditya Vyankatesh Mandke, Onkar Rupesh Litake and Dipali Dattatray Kadam	199
<i>An Extractive-Abstractive Approach for Multi-document Summarization of Scientific Articles for Literature Review</i>	
Kartik Shinde, Trinita Roy and Tirthankar Ghosal	204
<i>Overview of the DAGPap22 Shared Task on Detecting Automatically Generated Scientific Papers</i>	
Yury Kashnitsky, Drahomira Herrmannova, Anita de Waard, George Tsatsaronis, Catriona Catriona Fennell and Cyril Labbe	210
<i>SynSciPass: detecting appropriate uses of scientific text generation</i>	
Domenic Rosati	214
<i>Detecting generated scientific papers using an ensemble of transformer models</i>	
Anna Glazkova and Maksim Glazkov	223
<i>Overview of the SV-Ident 2022 Shared Task on Survey Variable Identification in Social Science Publications</i>	
Tornike Tsereteli, Yavuz Selim Kartal, Simone Paolo Ponzetto, Andrea Zielinski, Kai Eckert and Philipp Mayr	229
<i>Varanalysis@SV-Ident 2022: Variable Detection and Disambiguation Based on Semantic Similarity</i>	
Alica Hövelmeyer and Yavuz Selim Kartal	247
<i>Benchmark for Research Theme Classification of Scholarly Documents</i>	
Óscar E. Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi and Allan Hanbury	253
<i>Overview of the First Shared Task on Multi Perspective Scientific Document Summarization (MuP)</i>	
Arman Cohan, Guy Feigenblat, Tirthankar Ghosal and Michal Shmueli-Scheuer	263

<i>Multi Perspective Scientific Document Summarization With Graph Attention Networks (GATS)</i> Abbas Akkasi	268
<i>GUIR @ MuP 2022: Towards Generating Topic-aware Multi-perspective Summaries for Scientific Documents</i> Sajad Sotudeh and Nazli Goharian	273
<i>LTRC @MuP 2022: Multi-Perspective Scientific Document Summarization Using Pre-trained Generation Models</i> Ashok Urlana, Nirmal Surange and Manish Shrivastava	279
<i>Team AINLPML @ MuP in SDP 2021: Scientific Document Summarization by End-to-End Extractive and Abstractive Approach</i> Sandeep Kumar, Guneet Singh Kohli, Kartik Shinde and Asif Ekbal	285

