

EMNLP 2020

First Workshop on Scholarly Document Processing

Proceedings of the Workshop

November 19, 2020

Online

©2020 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-952148-70-5 (Volume 1)

Introduction

Welcome to the First Workshop on Scholarly Document Processing (SDP) at EMNLP 2020.

Next to keeping up with the growing literature in their own and related fields, scholars increasingly also need to rebut pseudo-science and disinformation. To address this challenge, computational work on enhancing search, summarization, and analysis of scholarly documents has flourished. However, the various strands of research on scholarly document processing remain fragmented. To reach to the broader NLP and AI/ML community, pool distributed efforts and enable shared access to published research, we held the 1st Workshop on Scholarly Document Processing at EMNLP20. The SDP workshop consisted of a research track and three Shared Tasks, geared towards easier access to scientific methods and results.

<https://ornlcda.github.io/SDProc/>

Organizers:

Muthu Kumar Chandrasekaran, Amazon, Seattle, USA Anita de Waard, Elsevier, USA Guy Feigenblat, IBM Research AI, Haifa Research Lab, Israel Dayne Freitag, SRI International, San Diego, USA Tirthankar Ghosal, Indian Institute of Technology Patna, India Eduard Hovy, Research Professor, LTI, Carnegie Mellon University, USA Petr Knoth, Open University, UK David Konopnicki, IBM Research AI, Haifa Research Lab, Israel Philipp Mayr, GESIS – Leibniz Institute for the Social Sciences, Germany Robert M. Patton, Oak Ridge National Laboratory, USA Michal Shmueli-Scheuer, IBM Research AI, Haifa Research Lab, Israel

Program Committee:

Please find our programme committee in the following webpage:

<https://ornlcda.github.io/SDProc/programcommittee.html>

Invited Speaker:

Kuansan Wang, Managing Director, MSR Outreach Academic Services, USA Steinn Sigurðsson, Scientific Director of arXiv, Professor in the Department of Astronomy & Astrophysics at The Pennsylvania State University

Table of Contents

<i>Overview of the First Workshop on Scholarly Document Processing (SDP)</i> Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer and Anita de Waard	1
<i>The future of arXiv and knowledge discovery in open science</i> Steinn Sigurdsson	7
<i>Acknowledgement Entity Recognition in COVID-19 Papers</i> Jian Wu, Pei Wang, Xin Wei, Sarah Rajtmajer, C Lee Giles and Christopher Griffin	10
<i>A Smart System to Generate and Validate Question Answer Pairs for COVID-19 Literature</i> Rohan Bhambhoria, Luna Feng, Dawn Sepehr, John Chen, Conner Cowling, Sedef Kocak and Elham Dolatabadi	20
<i>Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset</i> Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang and Jimmy Lin	31
<i>The impact of preprint servers in the formation of novel ideas</i> Swarup Satish, Zonghai Yao, Andrew Drozdov and Boris Veytsman	42
<i>Effective distributed representations for academic expert search</i> Mark Berger, Jakub Zavrel and Paul Groth	56
<i>Learning CNF Blocking for Large-scale Author Name Disambiguation</i> Kunho Kim, Athar Sefid and C Lee Giles	72
<i>Reconstructing Manual Information Extraction with DB-to-Document Backprojection: Experiments in the Life Science Domain</i> Mark-Christoph Müller, Sucheta Ghosh, Maja Rey, Ulrike Wittig, Wolfgang Müller and Michael Strube	81
<i>DeepPaperComposer: A Simple Solution for Training Data Preparation for Parsing Research Papers</i> Meng Ling and Jian Chen	91
<i>Improved Local Citation Recommendation Based on Context Enhanced with Global Information</i> Zoran Medić and Jan Snajder	97
<i>On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining</i> Ibrahim Burak Ozyurt	104
<i>SciWING— A Software Toolkit for Scientific Document Processing</i> Abhinav Ramesh Kashyap and Min-Yen Kan	113
<i>Multi-task Peer-Review Score Prediction</i> Jiyi Li, Ayaka Sato, Kazuya Shimura and Fumiyo Fukumoto	121
<i>ERLKG: Entity Representation Learning and Knowledge Graph based association analysis of COVID-19 through mining of unstructured biomedical corpora</i> Sayantan Basu, Sinchani Chakraborty, Atif Hassan, Sana Siddique and Ashish Anand	127

<i>Towards Grounding of Formulae</i>	
Takuto Asakura, André Greiner-Petter, Akiko Aizawa and Yusuke Miyao	138
<i>SChuBERT: Scholarly Document Chunks with BERT-encoding boost Citation Count Prediction.</i>	
Thomas van Dongen, Gideon Maillette de Buy Wenniger and Lambert Schomaker	148
<i>Structure-Tags Improve Text Classification for Scholarly Document Quality Prediction</i>	
Gideon Maillette de Buy Wenniger, Thomas van Dongen, Eleri Aedmaa, Herbert Teun Kruitbosch, Edwin A. Valentijn and Lambert Schomaker	158
<i>Cydex: Neural Search Infrastructure for the Scholarly Literature</i>	
Shane Ding, Edwin Zhang and Jimmy Lin	168
<i>On the Use of Web Search to Improve Scientific Collections</i>	
Krutarth Patel, Cornelia Caragea and Sujatha Das Gollapalli	174
<i>Scaling Systematic Literature Reviews with Machine Learning Pipelines</i>	
Seraphina Goldfarb-Tarrant, Alexander Robertson, Jasmina Lazic, Theodora Tsouloufi, Louise Donnison and Karen Smyth	184
<i>Document-Level Definition Detection in Scholarly Documents: Existing Models, Error Analyses, and Future Directions</i>	
Dongyeop Kang, Andrew Head, Risham Sidhu, Kyle Lo, Daniel Weld and Marti A. Hearst ...	196
<i>A New Neural Search and Insights Platform for Navigating and Organizing AI Research</i>	
Marzieh Fadaee, Olga Gureenkova, Fernando Rejon Barrera, Carsten Schnober, Wouter Weerkamp and Jakub Zavrel	207
<i>Overview and Insights from the Shared Tasks at Scholarly Document Processing 2020: CL-SciSumm, LaySumm and LongSumm</i>	
Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer and Anita de Waard	214
<i>CIST@CL-SciSumm 2020, LongSumm 2020: Automatic Scientific Document Summarization</i>	
Lei Li, Yang Xie, Wei Liu, Yinan Liu, Yafei Jiang, Siya Qi and Xingyuan Li	225
<i>NLP-PINGAN-TECH @ CL-SciSumm 2020</i>	
Ling Chai, Guizhen Fu and Yuan Ni	235
<i>IITBH-IITP@CL-SciSumm20, CL-LaySumm20, LongSumm20</i>	
Saichethan Reddy, NAVEEN SAINI, Sriparna Saha and Pushpak Bhattacharyya	242
<i>AUTH @ CLSciSumm 20, LaySumm 20, LongSumm 20</i>	
Alexios Gidiotis, Stefanos Stefanidis and Grigorios Tsoumakas	251
<i>UniHD@CL-SciSumm 2020: Citation Extraction as Search</i>	
Dennis Aumiller, satya almasian, Philip Hausner and Michael Gertz	261
<i>IITP-AI-NLP-ML@ CL-SciSumm 2020, CL-LaySumm 2020, LongSumm 2020</i>	
Santosh Kumar Mishra, Harshavardhan Kunderapu, Naveen Saini, Sriparna Saha and Pushpak Bhattacharyya	270
<i>IA-Team / Martin-Luther-Universität Halle-Wittenberg@CLSciSumm 20</i>	
Artur Jurk, Maik Boltze, Georg Keller, Lorna Ulbrich and Anja Fischer	277

<i>Team MLU@CL-SciSumm20: Methods for Computational Linguistics Scientific Citation Linkage</i> Rong Huang and Kseniia Krylova	282
<i>IR&TM-NJUST@CLSciSumm 20</i> Heng Zhang, Lifan Liu, Ruping Wang, Shaohu Hu, Shutian Ma and Chengzhi Zhang.....	288
<i>CiteQA@CLSciSumm 2020</i> Anjana Umapathy, Karthik Radhakrishnan, Kinjal Jain and Rahul Singh	297
<i>Dimsum @LaySumm 20</i> Tiezheng Yu, Dan Su, Wenliang Dai and Pascale Fung	303
<i>ARTU / TU Wien and Artificial Researcher@ LongSumm 20</i> Alaa El-Ebshihy, Annisa Maulida Ningtyas, Linda Andersson, Florina Piroi and Andreas Rauber 310	
<i>Monash-Summ@LongSumm 20 SciSummPip: An Unsupervised Scientific Paper Summarization Pipeline</i> Jiaxin Ju, Ming Liu, Longxiang Gao and Shirui Pan.....	318
<i>Using Pre-Trained Transformer for Better Lay Summarization</i> Seungwon Kim.....	328
<i>Summaformers @ LaySumm 20, LongSumm 20</i> Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta and Vasudeva Varma	336
<i>Divide and Conquer: From Complexity to Simplicity for Lay Summarization</i> Rochana Chaturvedi, SAACHI ., Jaspreet Singh Dhani, Anurag Joshi, Ankush Khanna, Neha Tomar, Swagata Duari, Alka Khurana and Vasudha Bhatnagar	344
<i>GUIR @ LongSumm 2020: Learning to Generate Long Summaries from Scientific Documents</i> Sajad Sotudeh Gharebagh, Arman Cohan and Nazli Goharian.....	356

Conference Program

Overview of the First Workshop on Scholarly Document Processing (SDP)

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer and Anita de Waard

The future of arXiv and knowledge discovery in open science

Steinn Sigurdsson

Research Track Session 1

Acknowledgement Entity Recognition in COVID-19 Papers

Jian Wu, Pei Wang, Xin Wei, Sarah Rajtmajer, C Lee Giles and Christopher Griffin

A Smart System to Generate and Validate Question Answer Pairs for COVID-19 Literature

Rohan Bhambhoria, Luna Feng, Dawn Sepehr, John Chen, Conner Cowling, Sedef Kocak and Elham Dolatabadi

Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset

Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang and Jimmy Lin

The impact of preprint servers in the formation of novel ideas

Swarup Satish, Zonghai Yao, Andrew Drozdov and Boris Veytsman

Research Track Session 2

Effective distributed representations for academic expert search

Mark Berger, Jakub Zavrel and Paul Groth

Learning CNF Blocking for Large-scale Author Name Disambiguation

Kunho Kim, Athar Sefid and C Lee Giles

Reconstructing Manual Information Extraction with DB-to-Document Backprojection: Experiments in the Life Science Domain

Mark-Christoph Müller, Sucheta Ghosh, Maja Rey, Ulrike Wittig, Wolfgang Müller and Michael Strube

No Day Set (continued)

Research Track Session 3

DeepPaperComposer: A Simple Solution for Training Data Preparation for Parsing Research Papers

Meng Ling and Jian Chen

Improved Local Citation Recommendation Based on Context Enhanced with Global Information

Zoran Medić and Jan Snajder

Poster Session

On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining

Ibrahim Burak Ozyurt

SciWING– A Software Toolkit for Scientific Document Processing

Abhinav Ramesh Kashyap and Min-Yen Kan

Multi-task Peer-Review Score Prediction

Jiyi Li, Ayaka Sato, Kazuya Shimura and Fumiyo Fukumoto

ERLKG: Entity Representation Learning and Knowledge Graph based association analysis of COVID-19 through mining of unstructured biomedical corpora

Sayantan Basu, Sinchani Chakraborty, Atif Hassan, Sana Siddique and Ashish Anand

Towards Grounding of Formulae

Takuto Asakura, André Greiner-Petter, Akiko Aizawa and Yusuke Miyao

SChuBERT: Scholarly Document Chunks with BERT-encoding boost Citation Count Prediction.

Thomas van Dongen, Gideon Maillette de Buy Wenniger and Lambert Schomaker

Structure-Tags Improve Text Classification for Scholarly Document Quality Prediction

Gideon Maillette de Buy Wenniger, Thomas van Dongen, Eleri Aedmaa, Herbert Teun Kruitbosch, Edwin A. Valentijn and Lambert Schomaker

Cydex: Neural Search Infrastructure for the Scholarly Literature

Shane Ding, Edwin Zhang and Jimmy Lin

No Day Set (continued)

On the Use of Web Search to Improve Scientific Collections

Krutarth Patel, Cornelia Caragea and Sujatha Das Gollapalli

Scaling Systematic Literature Reviews with Machine Learning Pipelines

Seraphina Goldfarb-Tarrant, Alexander Robertson, Jasmina Lazic, Theodora Tsouloufi, Louise Donnison and Karen Smyth

Document-Level Definition Detection in Scholarly Documents: Existing Models, Error Analyses, and Future Directions

Dongyeop Kang, Andrew Head, Risham Sidhu, Kyle Lo, Daniel Weld and Marti A. Hearst

A New Neural Search and Insights Platform for Navigating and Organizing AI Research

Marzieh Fadaee, Olga Gureenkova, Fernando Rejon Barrera, Carsten Schnober, Wouter Weerkamp and Jakub Zavrel

Shared Task Track

Overview and Insights from the Shared Tasks at Scholarly Document Processing 2020: CL-SciSumm, LaySumm and LongSumm

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer and Anita de Waard

CIST@CL-SciSumm 2020, LongSumm 2020: Automatic Scientific Document Summarization

Lei Li, Yang Xie, Wei Liu, Yinan Liu, Yafei Jiang, Siya Qi and Xingyuan Li

NLP-PINGAN-TECH @ CL-SciSumm 2020

Ling Chai, Guizhen Fu and Yuan Ni

IITBH-IITP@CL-SciSumm20, CL-LaySumm20, LongSumm20

Saichethan Reddy, NAVEEN SAINI, Sriparna Saha and Pushpak Bhattacharyya

AUTH @ CLSciSumm 20, LaySumm 20, LongSumm 20

Alexios Gidiotis, Stefanos Stefanidis and Grigorios Tsoumakas

UniHD@CL-SciSumm 2020: Citation Extraction as Search

Dennis Aumiller, satya almasian, Philip Hausner and Michael Gertz

IITP-AI-NLP-ML@ CL-SciSumm 2020, CL-LaySumm 2020, LongSumm 2020

Santosh Kumar Mishra, Harshavardhan Kunderapu, Naveen Saini, Sriparna Saha and Pushpak Bhattacharyya

No Day Set (continued)

IA-Team / Martin-Luther-Universität Halle-Wittenberg@CLSciSumm 20

Artur Jurk, Maik Boltze, Georg Keller, Lorna Ulbrich and Anja Fischer

Team MLU@CL-SciSumm20: Methods for Computational Linguistics Scientific Citation Linkage

Rong Huang and Kseniia Krylova

IR&TM-NJUST@CLSciSumm 20

Heng Zhang, Lifan Liu, Ruping Wang, Shaohu Hu, Shutian Ma and Chengzhi Zhang

CiteQA@CLSciSumm 2020

Anjana Umopathy, Karthik Radhakrishnan, Kinjal Jain and Rahul Singh

Dimsum @LaySumm 20

Tiezheng Yu, Dan Su, Wenliang Dai and Pascale Fung

ARTU / TU Wien and Artificial Researcher@ LongSumm 20

Alaa El-Ebshihy, Annisa Maulida Ningtyas, Linda Andersson, Florina Piroi and Andreas Rauber

Monash-Summ@LongSumm 20 SciSummPip: An Unsupervised Scientific Paper Summarization Pipeline

Jiaxin Ju, Ming Liu, Longxiang Gao and Shirui Pan

Using Pre-Trained Transformer for Better Lay Summarization

Seungwon Kim

Summaformers @ LaySumm 20, LongSumm 20

Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta and Vasudeva Varma

Divide and Conquer: From Complexity to Simplicity for Lay Summarization

Rochana Chaturvedi, SAACHI ., Jaspreet Singh Dhani, Anurag Joshi, Ankush Khanna, Neha Tomar, Swagata Duari, Alka Khurana and Vasudha Bhatnagar

GUIR @ LongSumm 2020: Learning to Generate Long Summaries from Scientific Documents

Sajad Sotudeh Gharebagh, Arman Cohan and Nazli Goharian