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 Melon 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

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 CORD-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)

Shared Task Track

No Day Set (continued)

1A-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 Umapathy, 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