First Workshop on Scholarly Document Processing (SDP)

Online Workshop @EMNLP 2020
November 19, 2020
https://ornlcda.github.io/SDProc/

@SDProc #sdpws2020

Introduction to the workshop: Motivation

The general research **community on scholarly document processing is fragmented** (SIGIR, JCDL, LREC, … ACL, EMNLP, ….)

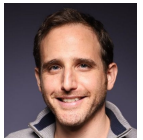
We offer an **interdisciplinary venue** for researchers interested in any aspect of **mining scientific literature**

Workshop is focused on **enhancing search, retrieval, summarization, and analysis of scholarly documents**

Overview paper: https://www.aclweb.org/anthology/2020.sdp-1.1.pdf

Workshop Team







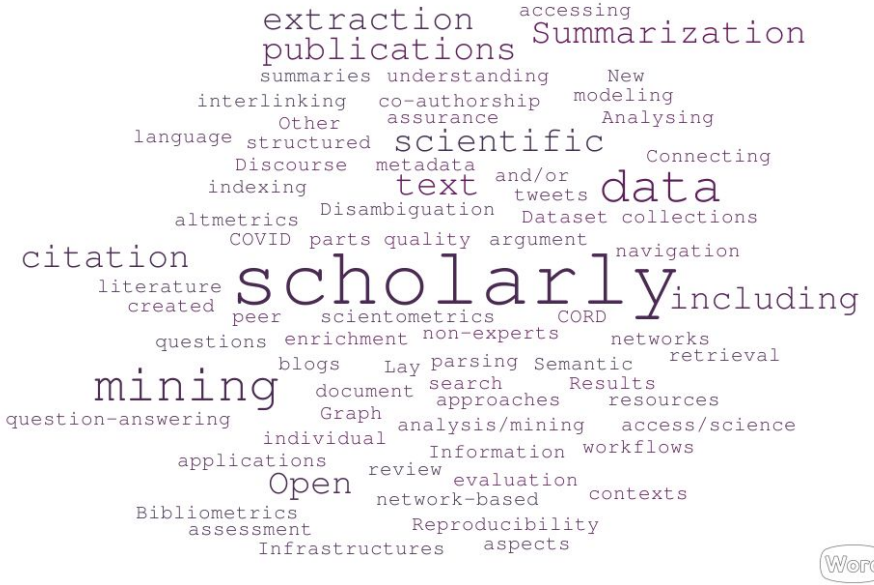
Previous workshops mainly at JCDL, SIGIR conference:
Joint Workshop on Bibliometric-enhanced IR and NLP for Digital Libraries (**BIRNDL**) + International Workshop on Mining Scientific Publications (**WOSP**)

Organizers: Muthu Kumar **Chandrasekaran** (Amazon, USA), Anita **de Waard** (Elsevier, USA), Guy **Feigenblat** (IBM Research, Israel), Dayne **Freitag** (SRI International, USA), Tirthankar **Ghosal** (IIT Patna, India), Eduard **Hovy** (Carnegie Melon University, USA), Petr Knoth (Open University, UK), David Konopnicki (IBM Research, Israel), Philipp **Mayr** (GESIS, Germany), Robert M. Patton (Ridge National Laboratory, USA), Michal **Shmueli-Scheuer** (IBM Research, Israel), Dominika Tkaczyk (Crossref, UK) → https://ornlcda.github.io/SDProc/organizingcommittee.html

Topics of SDP CfP

All approaches to mining scholarly data

- Information extraction, text mining and parsing of **scholarly literature**
- Reproducibility and peer review
- Lay **Summarization** (i.e., summaries created for non-experts) of individual and collections of scholarly documents
- Discourse modeling and argument mining
- Summarization and question-answering for scholarly documents
- Semantic and network-based indexing, search and navigation in structured text
- Graph analysis/mining including citation and co-authorship networks
- Analysing and mining of citation contexts for document understanding and retrieval
- New scholarly language resources and evaluation
- Connecting and interlinking publications, data, tweets, blogs or their parts
- Disambiguation, metadata **extraction**, enrichment, and data quality assurance for scholarly documents
- Bibliometrics, scientometrics, and altmetrics approaches and applications
- Other aspects of scientific workflows including open access/science, and research assessment
- Infrastructures for accessing scientific publications and/or research data
- Results and research questions on the **COVID-19 Open Research** Dataset (CORD-19)



Submissions to SDP 2020

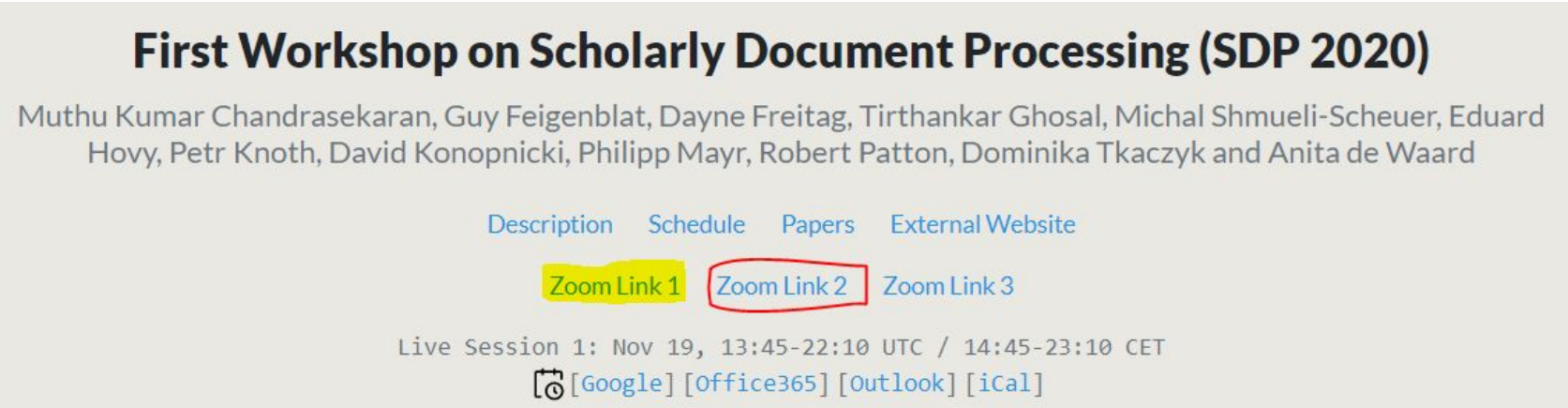
34 papers were submitted to the **research track**

- **7** accepted as full papers + **2** as short papers
- **14** papers rejected
- **11** papers + **1** “Findings of EMNLP” paper were invited as posters
- **3** “Findings of EMNLP” papers were invited to the short paper session
- **1** demo paper as technical paper in the poster session

18 shared task papers we submitted to the **shared task track**

SDP Proceedings: https://www.aclweb.org/anthology/volumes/2020.sdp-1/ (361 pages)

Agenda of SDP



First Workshop on Scholarly Document Processing (SDP 2020)

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Michal Shmueli-Scheuer, Eduard Hovy, Petr Knoth, David Konopnicki, Philipp Mayr, Robert Patton, Dominika Tkaczyk and Anita de Waard

Description Schedule Papers External Website

Zoom Link 1 Zoom Link 2 Zoom Link 3

Live Session 1: Nov 19, 13:45-22:10 UTC / 14:45-23:10 CET

[Google] [Office365] [Outlook] [iCal]

- **Research Track**: 3 sessions, 2 keynotes and poster pitches → **Zoom Link 1**
- **Shared Task Track**: all presentations → **Zoom Link 2**
- **Joint Poster Session**: will be in Gather.Town, Room **N**

All pre-recorded talks are linked here:
https://ornlcda.github.io/SDProc/program.html

Agenda of SDP

Time in EST	Plenary -- Zoom Link 1	Breakout -- Zoom Link 2	Presentations in Shared Task Track	
8:45	*Workshop Start*			
8.45-9:00	Introduction to SDP			
9:00-9:15	Teaser for Shared Tasks (5 mins each)	Shared Task: 9:15-10:40		
9:15 - 10:35	*Research Track: Session 1 COVID-19 Document Processing*	Shared Task 1: 9:15-9:55	9:15-9:27	CL-SciSumm (Task 1a): Chai et al., NLP-PINGAN-TECH @ CL-SciSumm 2020
9:15-9:35	Wu et al., Acknowledgement Entity Recognition in CORD-19 Papers		9:28-9:40	LaySumm 2: Yu et al., Dimsum @LaySumm 20
9:35-9:55	Bhambhoria et al., A Smart System to Generate and Validate Question Answer Pairs for COVID-19 Literature.		9:40-9:55	LaySumm 3 & LongSumm 2: Ghosh Roy et al., Summaformers @ LaySumm 20, LongSumm 20
9:55-10:15	Zhang et al., Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset.	Shared Task 2: 9:55-10:40	9:55-10:10	Li et al., CIST@CL-SciSumm 2020, LongSumm 2020: Automatic Scientific Document Summarization
10:15-10:35	Satish et al., The Impact of Preprint Servers in the Formation of Novel Ideas.		10:10-10:25	LongSumm 3 & CL-SciSumm (Task 2): Reddy et al., IIITBH-IITP@CL-SciSumm20, CL-LaySumm20, LongSumm20
			10:25-10:40	Gidiotis et al., AUTH @ CLSciSumm 20, LaySumm 20, LongSumm 20
10:40-11:25	**Keynote 1: Kuansan Wang** **Mitigating Scholarly Corpus Biases with Citations: A Case Study on CORD-19**			

Plenary

Parallel

Agenda of SDP

	Research Track: Session 2 SDP Mixed Session	*Shared Task Track Continues*		
11:50-12:10	Berger et al., Effective Distributed Representations for Academic Expert Search.	hared Task : 1:50 - 12:50	11:50 - 12:05	CL-SciSumm (Task 1a): Aumiller et al., UniHD@CL-SciSumm 2020: Citation Extraction as Search
12:10-12:30	Kim et al., Learning CNF Blocking for Large-scale Author Name Disambiguation.		12:05 - 12:17	LaySumm 1: Seungwon Kim, Using Pre-Trained Transformer for Better Lay Summarization
12:30-12:50	Müller et al., Reconstructing Manual Information Extraction with DB-to-Document Backprojection: Experiments in the Life Science Domain.		12:17 - 12:29	LongSumm 1: Gharebagh et al., GUIR @ LongSumm 2020: Learning to Generate Long Summaries from Scientific Documents
			12:30 - 12:42	Umapathy et al., CiteQA@CLSciSumm 2020
			12:42 - 12:50	Poster Session Announcement
12:50-13:30	*Poster Pitches*	*Poster Track*		
	11 Posters + 2 Demo Presentations	Posters: Gather Town		
13:30 - 14:00	*Break*			
	Research Track: Session 3: Short papers and Findings			
14:00-14:10	Ling & Chen, DeepPaperComposer: A Simple Solution for Training Data Preparation for Parsing Research Papers			
14:10 - 14:20	Medic & Snajder, Improved Local Citation Recommendation Based on Context Enhanced with Global Information			
14:20-14:30	Cao et al., Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora (**Findings of EMNLP**)			

Parallel

Plenary

Plenary for the rest of the day

Keynotes: live

Morning keynote at 10:40 EST

Kuansan Wang:
Managing Director, MSR Outreach Academic Services, USA

Mitigating scholarly corpus biases with citations: A case study on CORD-19

Afternoon keynote at 15:30 EST

Steinn Sigurðsson:
Scientific Director of arXiv, Professor in the Department of Astronomy & Astrophysics at The Pennsylvania State University

The future of arXiv and knowledge discovery in open science

Greeting Note by Eduard Hovy

Scholarly documents present several interesting challenges where DNNs have not nearly reached their potential:

- Scientific discourse: doc structure, not just a ‘flat’ event-sequence story
- Summarization: abstraction, not extraction

Is there a move away from surface-level info transformation toward semantics and content?

- Same task with multiple languages and different media
- Explainability
- DNNs as transformation engines AND knowledge bases

Defining structure and marrying transformation processes into that

Teaser for Shared Task Track

Teaser for Shared Tasks (5 mins each)

1. CL-SciSumm 2020
2. CL-LaySumm 2020
3. LongSumm 2020

Contact

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

Contact: sdproc@googlegroups.com

Follow us: https://twitter.com/SDProc

Get in touch with us to learn about **SDP 2021**!