Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019)

Muthu Kumar Chandrasekaran
SRI International, Menlo Park, USA
cmkumar087@gmail.com

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences, Germany
philipp.mayr@gesis.org

Michihiro Yasunaga
Yale University, USA
michi@yale.edu

Dayne Freitag
SRI International, San Diego, USA
freitag@ai.sri.com

Dragomir Radev
Yale University, USA
dragomir.radev@yale.edu

Min-Yen Kan
School of Computing, National University of Singapore, Singapore
kanmy@comp.nus.edu.sg

ABSTRACT

The deluge of scholarly publication poses a challenge for scholars find relevant research and policy makers to seek in-depth information and understand research impact. Information retrieval (IR), natural language processing (NLP) and bibliometrics could enhance scholarly search, retrieval and user experience, but their use in digital libraries is not widespread. To address this gap, we propose the 4^{th} Joint Workshop on BIRNDL and the 5^{th} CL-SciSumm Shared Task. We seek to foster collaboration among researchers in NLP, IR and Digital Libraries (DL), and to stimulate the development of new methods in NLP, IR, recommendation systems and scientometrics toward improved scholarly document understanding, analysis, and retrieval at scale.

CCS CONCEPTS

• **Information systems → Information retrieval**; *Link and co-citation analysis*; • **Applied computing → Digital libraries and archives**.

KEYWORDS

Scientometrics; Information Retrieval; Digital Libraries; NLP; Summarization; Information Extraction; Citation analysis

ACM Reference Format:
Muthu Kumar Chandrasekaran, Philipp Mayr, Michihiro Yasunaga, Dayne Freitag, Dragomir Radev, and Min-Yen Kan. 2019. Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3331184.3331650

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331650

1 INTRODUCTION

Over the past several years and at premier conferences, BIRNDL [2, 5, 6], together with its parent workshops, has established itself as the primary interdisciplinary venue for cross-pollination of NLP, IR and DL research. We observe that while some researchers are active in all three communities, the main discourse in these fields consist of different approaches to solve similar problems. A common forum for discussion benefits all communities, by catalyzing new ideas and collaborations and facilitates knowledge transfer. A recent description of the symbiotic relationship that exists among bibliometrics, IR and NLP was presented by Wolfram at the first BIRNDL workshop at JCDL 2016 [9].

The goal of the BIRNDL workshop at SIGIR 2019 is to engage the IR community in the open problems in Big Sciences. Big Science refers to the large, cross-domain digital repositories which index research papers, such as the ArXiv, ACM Digital Library, PubMed, ACL Anthology, IEEE database, Web of Science and Google Scholar. Currently, digital libraries collect and allow access to digital papers and their metadata—inclusive of citations—. However, they do not analyze and further utilize the items they index. The volume of scholarly publication poses a challenge for scholars in their search for relevant literature. Finding relevant scholarly literature is the key focus of the workshop and sets the agenda for tools and approaches to be discussed and evaluated at BIRNDL.

The 4^{th} BIRNDL workshop[1] and the 5^{th} CL-SciSumm Shared Task will be a follow-up to the previous editions co-located with SIGIR '18[2] where 6 research papers and 11 shared task system papers were presented[3] [6]. BIRNDL at SIGIR '19 has attracted 25 submissions to the research track and 16 registrations to the CL-SciSumm Shared Task track. It will also host two keynotes from acdemia and industry. BIRNDL at SIGIR '19 will be a full-day workshop.

As in the previous editions, the CL-SciSumm Shared Task generated a lot of interest and participation, and all participants strongly favored a follow-up this year. We have been regularly coordinating

[1] http://wing.comp.nus.edu.sg/~birndl2019/
[2] http://wing.comp.nus.edu.sg/birndl-sigir2018/
[3] http://ceur-ws.org/Vol-2132/

workshop series at premier IR and Information Systems venues—such as the Bibliometric-enhanced Information Retrieval (BIR) workshop series since 2014 [8] at ECIR and the NLPIR4DL workshop at ACL-IJCNLP, 2009.

Papers and talks at the workshop will incorporate insights from IR, NLP and bibliometrics to develop new techniques to address the open problems in scholarly digital libraries and studying impact of big science, such as evidence-based searching, measurement of research quality, relevance and impact, the emergence and decline of research problems, identification of scholarly relationships and influences and applied problems such as language translation, question-answering and summarization. We will also address the need for established, standardized baselines, evaluation metrics and standardised reference datasets. Towards the purpose of evaluating tools and technologies developed for digital libraries, we will organize the 5^{th} CL-SciSumm Shared Task– based on the CL-SciSumm corpus, comprising over 500 computational linguistics research papers, interlinked through a citation network. In this iteration of CL-SciSumm, we are increasing the size of our existing corpus by one-fold through automated annotations.

This workshop will be relevant to scholars in computer and information science, specialized in IR and NLP. It will also be of importance to all stakeholders in the publication pipeline: implementers, publishers and policymakers. Today's publishers continue to provide new ways to support their consumers in disseminating and retrieving the right published works to their readership. Formal citation metrics are increasingly a factor in decision-making by universities and funding bodies worldwide, making the need for research in applying these metrics more pressing.

2 WORKSHOP TOPICS AND FORMAT

By design, BIRNDL is an inclusive and diverse venue, in terms of both constituency and research. To promote a diverse constituency, we explicitly encourage female first authors. We invite stimulating research on topics including, but not limited to, full-text analysis, including multilingual analysis, IR methods for DL, and applications of citation-based NLP. Specific examples of fields of interest include:

– Infrastructure for scientific text mining and IR,
– Semantic and network-based indexing, search and navigation in structured text,
– Bibliometrics and citation analysis,
– Discourse modeling and argument mining,
– Summarization and question-answering for scholarly DLs,
– Recommendation for scholarly papers, reviewers, citations and publication venues,
– Measurement of document quality and impact,
– Information extraction and parsing tasks on scientific documents,
– Science knowledge base population (Sci-KBP) and inference,
– Automated discovery and maintenance of metadata and controlled vocabularies,
– Disambiguation and data quality assurance in scholarly DLs.

Importantly, to address the scarcity of validated datasets in this area, we also invite papers describing new and existing datasets. Submissions in this track will include instructions for accessing the data; metadata and documentation on its organization, content, and quality; and descriptions of possible use cases.

2.1 Tentative Schedule of Events

Along with research paper presentations, the workshop will host two keynotes. Prof. Bonnie Webber (University of Edinburgh) will deliver the Distinguished Keynote while Alex Wade from the Chan-Zuckerberg Initiative (CZI)'s *Meta*, will deliver a second keynote on "Personalized feed/query-formulation, predictive impact, and ranking". The schedule will include an overview of Shared Task design and results, followed by presentations from selected participants, with a poster session to accommodate additional Shared Task and research participants. The workshop will end with planning and discussion to decide on future directions and improvements to the workshop and the Shared Task.

2.2 The CL-SciSumm Shared Task

CL-SciSumm is the first medium-scale (over 500 annotated documents as of today) shared task on scientific document summarization in the computational linguistics domain. The 5^{th} CL-SciSumm[4] follows up on the successful previous editions conducted as a part of the BIRNDL workshops since 2016 and as a Pilot Task at the Text Analysis Conference 2014 (TAC 2014). This task is expected to be of interest to a broad community, including those working in CL and NLP, especially in the sub-disciplines of text summarization, discourse structure in scholarly discourse, paraphrase, textual entailment and text simplification. Since the CL-SciSumm '18, we have started working with collaborators at Yale University and enriched our dataset with metadata from the ACL Anthology Network (AAN)[5], doubling the size of our corpus.

For CL-SciSumm '19 we encourage NLP research groups to field systems using deep learning and other recent but data hungry methods to push the state-of-the-art. To this end, we are introducing a new dataset which expands the annotated CL-SciSumm corpus [10]. Further we are also introducing distantly supervised corpus of several thousands citance-reference sentence pairs to train systems for Task 1a.

In the CL-SciSumm 2018 Shared Task, fourteen teams registered, out of which eleven teams submitted their systems for evaluation and presented their results at BIRNDL [6]. Sixteen teams have CL-SciSumm 2019 and are working towards submissions. The Shared Task results will be published in a overview paper as part of the BIRNDL '19 proceedings.

The Shared Task comprises three sub-tasks in automatic scholarly document summarization on a corpus of research papers. Systems are provided with a Reference Paper (RP) and 10 or more Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (i.e., citances) that pertain to a particular citation to the RP are annotated. Systems are tasked to identify text spans in the RP that citances in CP are referring to (Task 1a); then assign one of the four discourse facets to identified reference span (Task 1b); and finally output a summary of the RP using outputs from Task 1 and other text extracted from RP.

[4] http://wing.comp.nus.edu.sg/~cl-scisumm2019/
[5] http://clair.eecs.umich.edu/aan/index.php

3 RELATED WORKSHOPS

Our workshop is a continuation of several previous ones on similar topics. We present a summary of some relevant recent events, which underpin our claim of the workshop topic being spot-on and relevant.

The following related workshops (NLPIR4DL, BIR, CLBib and the CL Summarization Pilot Task) have been organized by the BIRNDL organisers.

- 1^{st} BIRNDL at JCDL, 2^{nd} and 3^{rd} BIRNDL at SIGIR '17, Tokyo and '18, Ann Arbor,
- 1^{st} Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL) was held in conjunction with ACL-IJCNLP 2009, Singapore. It comprised 11 full papers (acceptance rate: 21%).
- 8^{th} Workshop on Bibliometric-enhanced Information Retrieval (BIR2019) at ECIR 2019. The focus of the BIR workshops at ECIR (2014, 2015, 2016, 2017, 2018 and 2019) was on research papers in information retrieval, information seeking, science modelling, network analysis, and digital libraries, applying insights from bibliometrics, scientometrics, and informetrics.
- 2^{nd} Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics (CLBib) at ISSI 2017 brought together researchers to study the ways Bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and NLP.
- The Computational Linguistics Pilot Task, held as a part of the Biomedical Summarization track, at TAC 2014 [4], where the results from 3 system papers were presented.

4 OUTLOOK

BIRNDL is an effort to bring together all interested parties working in the digital library space. It has always been an interdisciplinary venue to benefit Bibliometrics, IR and NLP. Previous editions of BIRNDL have created a greater impact than originally thought. BIRNDL 2016, led to a special issue on "Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries" with thirteen full journal papers, one of the largest special issues in the International Journal on Digital Libraries (IJDL) [7]. Another special issue on "Bibliometric-enhanced Information retrieval and Scientometrics" was published in *Scientometrics* in 2018 [3]. Recently, the Open Access journal *Frontiers in Research Metrics and Analytics* published a Research Topic with seven research articles on "Mining Scientific Papers: NLP-enhanced Bibliometrics" [1].

BIRNDL has a global presence with program committee memberships, submissions and participation across Europe, Asia, Australia and the Americas; with scholarship from computer science and social sciences; with scholarship from academia and industry. For a future edition, we call upon the leaders in the field, to unite under a common umbrella conference that represents the interests of all the communities in the digital library and bibliometrics space.

Selected BIRNDL papers will be invited to a special issue on "Mining Knowledge from Scientific Data" in the journal *Expert Systems*. All other presenters are invited for an other special issue in *Scientometrics*.

Since 2016 we have been maintaining the "Bibliometric-enhanced-IR Bibliography" [6] which is a bibliography of all scientific papers that appear in BIR, BIRNDL and related workshops and journal special issues.

ACKNOWLEDGMENTS

We thank SRI International and Chan-Zuckerberg Initiative (CZI) for their generous support in funding the organization of the CL-SciSumm Shared Task 2019. CZI sponsored Alex Wade's keynote. We immensely than Prof. Dragomir Radev and Michihiro Yasunaga from Yale University for sharing the SciSummNet dataset for CL-SciSumm 2019 and co-organising this time. This work by Philipp Mayr was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant number MA 3964/10-1, the "Establishing Contextual Dataset Retrieval - transferring concepts from document to dataset retrieval (ConDATA)" project.

We are also grateful to the co-organizers of the 1^{st} BIRNDL workshop - Guillaume Cabanac, Ingo Frommholz, and Dietmar Wolfram, for their continued support and involvement.

REFERENCES

[1] Iana Atanassova, Marc Bertin, and Philipp Mayr. 2019. Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. *Frontiers in Research Metrics and Analytics* (2019). https://doi.org/10.3389/frma.2019.00002

[2] Guillaume Cabanac, Muthu Kumar Chandrasekaran, Ingo Frommholz, Kokil Jaidka, Min-Yen Kan, Philipp Mayr, and Dietmar Wolfram. 2016. Report on the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). *SIGIR Forum* 50, 2 (2016), 36–43. http://sigir.org/wp-content/uploads/2017/01/p036.pdf

[3] Guillaume Cabanac, Ingo Frommholz, and Philipp Mayr. 2018. Bibliometric-enhanced information retrieval: preface. *Scientometrics* 116, 2 (2018), 1225–1227. https://doi.org/10.1007/s11192-018-2861-0

[4] Kokil Jaidka, Muthu Kumar Chandrasekaran, Beatriz Fisas Elizalde, Rahul Jha, Christopher Jones, Min-Yen Kan, Ankur Khanna, Diego Molla-Aliod, Dragomir R Radev, Francesco Ronzano, et al. 2014. The computational linguistics summarization pilot task. In *Proceedings of Text Analysis Conference*. Gaithersburg, USA.

[5] Philipp Mayr, Muthu Kumar Chandrasekaran, and Kokil Jaidka. 2017. Report on the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017). *SIGIR Forum* 51, 2 (2017), 107–113. http://sigir.org/wp-content/uploads/2018/01/p107.pdf

[6] Philipp Mayr, Muthu Kumar Chandrasekaran, and Kokil Jaidka. 2018. Report on the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018). *SIGIR Forum* 52, 2 (2018), 105–110. http://sigir.org/wp-content/uploads/2019/01/p105.pdf

[7] Philipp Mayr, Ingo Frommholz, Guillaume Cabanac, Muthu Kumar Chandrasekaran, Kokil Jaidka, Min-Yen Kan, and Dietmar Wolfram. 2018. Introduction to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). *International Journal on Digital Libraries* 19, 2-3 (2018), 107–111. https://doi.org/10.1007/s00799-017-0230-x

[8] Philipp Mayr, Andrea Scharnhorst, Birger Larsen, Philipp Schaer, and Peter Mutschke. 2014. Bibliometric-Enhanced Information Retrieval. In *36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, Maarten et al. de Rijke (Ed.). Springer International Publishing, Amsterdam, The Netherlands, 798–801. https://doi.org/10.1007/978-3-319-06028-6_99

[9] Dietmar Wolfram. 2016. Bibliometrics, Information Retrieval and Natural Language Processing: Natural Synergies to Support Digital Library Research. In *Proc. of the BIRNDL Workshop 2016*. 6–13. http://ceur-ws.org/Vol-1610/paper1.pdf

[10] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri Irene Li Dan, and Friedman Dragomir R Radev. 2019. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. (2019).

[6] https://github.com/PhilippMayr/Bibliometric-enhanced-IR_Bibliography/