Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018)

Muthu Kumar Chandrasekaran[1], Kokil Jaidka[2], Philipp Mayr[3]
[1]NUS School of Computing, Singapore
[2]University of Pennsylvania, USA
[3]GESIS – Leibniz Institute for the Social Sciences, Germany
[1]muthu.chandra@comp.nus.edu.sg; [2]jaidka@sas.upenn.edu; [3]philipp.mayr@gesis.org

ABSTRACT

The large scale of scholarly publications poses a challenge for scholars in information seeking and sensemaking. Information retrieval (IR), bibliometric and NLP techniques could help in these search and look-up activities, but are not yet widely used. To this purpose, we propose the third iteration of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) [3, 1]. The workshop is intended to stimulate IR researchers and Digital Library professionals to elaborate on new approaches in natural language processing, information retrieval, scientometrics, text mining and recommendation techniques that can advance the state-of-the-art in scholarly document understanding, analysis, and retrieval at scale. The BIRNDL workshop will incorporate multiple invited talks, paper sessions, a poster session and the third edition of the *Computational Linguistics (CL) Scientific Summarization Shared Task*.

CCS Concepts

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

Keywords

Bibliometrics; Information Retrieval; Digital Libraries; Natural Language Processing; Text Mining; Information Extraction; Citation analysis

1. INTRODUCTION

Over the past many years and at major conferences for information retrieval (IR) research, the BIRNDL workshops [3, 1] and its parent workshops are establishing themselves as the primary interdisciplinary venue for the cross-pollination of natural language processing (NLP), IR and bibliometrics. The workshop series is motivated by the observation that while the membership in either community shares only a partial overlap; yet, the main discourse in both fields consists of different approaches to solve similar problems. A common forum for discussion benefits both communities, by catalyzing new ideas and collaborations and facilitating knowledge transfer. A recent description of the symbiotic relationship that exists among bibliometrics, IR and NLP has been presented by Wolfram [8]. The highlights of SIGIR-17's BIRNDL workshop were published in SIGIR Forum [3].

The goal of the BIRNDL workshop at SIGIR 2018 is to engage the IR community about the open problems in Big Sciences. Big Science refers to the large, cross-domain digital repositories which index research papers, such as the ACL Anthology, ArXiv, ACM Digital Library, PubMed, IEEE database, Web of Science and Google Scholar. Currently, digital libraries collect and allow access to digital papers and their metadata—inclusive of citations—but mostly do not analyze the items they index. The scale of scholarly publications 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 3^{rd} BIRNDL workshop and 4^{th} CL-SciSumm Shared Task will be a follow-up to the 2^{nd} BIRNDL workshop and 3^{rd} CL-SciSumm Shared Task, co-located with SIGIR 2017[1], where 7 research papers and 9 system papers were presented[2] [3]. The keynote by Simone Teufel (University of Cambridge, UK) discussed how citation links and entailment can be applied to answer global scientometric questions [7]. The Shared Task generated a lot of interest and participation, and all proponents strongly favored a follow-up this year. The main organizers have regularly been coordinating workshop series at premier IR and IS venues - such as the Bibliometric-enhanced Information Retrieval (BIR) workshops in 2014, 2015, 2016 and 2017 at ECIR [4] and the NLPIR4DL workshop at ACL-IJCNLP (2009). In 2018, the BIRNDL workshop plans to take this legacy forward with a special focus on scholarly publications and new datasets, and an updated scientific summarization Shared Task for its participants.

Papers and talks at the workshop will incorporate insights from IR, NLP and bibliometrics to develop new tech-

[1]http://wing.comp.nus.edu.sg/birndl-sigir2017/
[2]http://ceur-ws.org/Vol-1888/ and http://ceur-ws.org/Vol-2002/

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

niques to address the open problems in 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 test collections. Towards the purpose of evaluating tools and technologies developed for digital libraries, we will organize the 4th 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 adding to our existing organization team and also nearly doubling the size of our existing dataset.

This workshop will be relevant to scholars in computer and information science, specialized in IR and NLP. It will also be of importance for 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 audience. 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

Our goal is to encourage insights from IR, NLP and computational linguistics for scholarly document understanding, document analysis and retrieval in digital libraries. We invite stimulating submissions on topics including – but not limited to – full-text analysis, multimedia and multilingual analysis and alignment as well as the application of citation-based NLP, information retrieval and information seeking techniques in digital libraries. Specific examples of fields of interests include (but are not limited to):

- Infrastructures for scientific text mining and IR,
- Semantic and network-based indexing, navigation, searching and browsing in structured data,
- Information extraction and parsing tasks in scientific papers,
- Population of a science knowledge base and performing inference on it,
- Bibliometrics, citation analysis and network analysis for IR,
- Discourse structure identification and argument mining from scientific papers,
- Summarization and question-answering for scholarly DLs,
- Recommendation for scholarly papers, reviewers, citations and publication venues,
- Measurement and evaluation of quality and impact,
- Metadata and controlled vocabularies for resource description and discovery; automatic metadata discovery, such as language identification,
- Disambiguation issues in scholarly DLs using NLP or IR techniques; data cleaning and data quality.

Additionally, this year we also invite dataset papers which describe new and pre-existing data resources. This is to address the other challenge in bibliometrics research – the scarcity of validated datasets for problem solving, benchmarking and evaluation. Dataset paper submissions must comprise:

- The data itself – organized as a single dataset or a group of datasets, and
- Metadata which describes data collection and processing methods, documentation of the structure and descriptive statistics about the content and quality of the dataset.
- Authors should describe potential uses and applications of the dataset, but any sophisticated analysis can be a regular paper submission.

2.1 Tentative Schedule of Events

We plan to follow a similar schedule as our 2017 workshop at SIGIR. For 2018 we apply for a full day. After notification of acceptance at SIGIR, we will publish an open call for papers. All submissions will be reviewed by at least three independent reviewers from the Program Committee. The full day workshop will start with a keynote followed by regular research paper presentations in the forenoon. The afternoon sessions will also feature an invited talk, followed by an overview paper on the Shared Task and selected presentations of the participating teams in the Shared Task. In a poster session in the afternoon, a few other participants of the Shared Task and papers deemed more suited for a poster than a presentation will be invited to display a poster or demo their system. We will end the workshop with a planning and discussion session to decide on further directions and enhancements to the workshop and the Shared Task. We anticipate at least 50 attendees at our workshop. This is based on the attendance at the second BIRNDL workshop at SIGIR 2017.

2.2 The CL-SciSumm Shared Task

The 4^{th} Computational Linguistics (CL) Scientific Summarization Shared Task is sponsored by Microsoft Research Asia and will be conducted as a part of this workshop. This is the first medium-scale shared task on scientific document summarization in the computational linguistics domain. It follows up on and extends the corpus sizes of the successful CL Shared Task conducted as a part of the BIRNDL workshops in 2017, 2016 and the Pilot Task conducted as a part of the BiomedSumm Track at the Text Analysis Conference 2014 (TAC 2014) [2]. In the CL-SciSumm 2017 Shared Task, 15 teams signed up, and nine teams ultimately submitted and presented their results.

The Shared Task comprises three sub-tasks in automatic research paper summarization on a new corpus of research papers, as described below.

Given: A topic consisting of 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) have been identified that pertain to a particular citation to the RP.

- Task 1a: For each citance, identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. These are of the granularity of a sentence fragment, a full sentence, or several consecutive sentences (no more than 5).
- Task 1b: For each cited text span, identify what facet of the paper it belongs to, from a predefined set of facets.

- Task 2 (optional bonus task): Finally, generate a structured summary of the RP from the cited text spans of the RP. The length of the summary should not exceed 250 words. Evaluation: Task 1 will be scored by overlap of text spans measured by number of sentences in the system output vs gold standard. Task 2 will be scored using the ROUGE family of metrics between the system output, and i) human summaries, ii) community summaries comprising the cited text spans, and ii) the Abstract section of the reference paper.

The CL-SciSumm corpus comprises a training corpus of forty topics and a test corpus of ten topics. Each topic comprises ACL Computational Linguistics research papers, and their citing papers and three output summaries each. The three output summaries comprise: human summaries, faceted summaries of the traditional self-summary (the abstract) and the community summary (the collection of citation sentences or *citances*) [6]. For the 2018 Shared Task, our team is joining hands with the Language, Information and Learning at Yale (LILY) group at Yale University. We have also enriched our dataset by nearly doubling the size of our corpus and incorporating metadata from the ACL Anthology Network (AAN)[3]

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.

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 proposers.

- 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%).
- 5^{th} Workshop on Bibliometric-enhanced Information Retrieval (BIR2017) at ECIR 2017 [4]. The focus of the BIR workshops at ECIR (2014, 2015, 2016 and 2017) was on research papers in information retrieval, information seeking, science modelling, network analysis, and digital libraries, applying insights from bibliometrics, scientometrics, and informetrics.
- 1^{st} Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics (CLBib) at ISSI 2015 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 [2], where the results from 3 system papers were presented.

The following workshops have been organized by other research groups.

- 6^{th} Workshop on Mining Scientific Publications (WOSP) at JCDL 2017. This workshop series is co-located with the JCDL and tries to leverage the potential of text and data mining technologies to improve the process of how research is done.
- Scholarly Big Data: AI Perspectives, Challenges, and Ideas at AAAI 2016 and IJCAI 2016. This workshop series is related to our topics and reflects many of the same Program Committee members. It indicates a high degree of interest for our topic, and will be synergistic due to its complementary date.
- 3^{rd} Workshop on Argumentation Mining at ACL 2016. This related workshop is synergistic and complementary. We overlap to a small extent in being interested in argumentation (their workshop) in scientific documents (our workshop).

4. OUTLOOK

This workshop is a next step to foster a reflection on interdisciplinarity, and the benefits that the disciplines Bibliometrics, IR and NLP can derive from it in a digital libraries context. Continuing our tradition of producing follow-up special issues after our workshops, we will invite the authors of accepted papers at this year's BIRNDL workshop to submit extended versions to a special issue in a highly visible and prestigious journal. As an output of BIRNDL 2016, we have published a special issue on "Bibliometrics, Information Retrieval and Natural Language Processing in Digital Libraries" in the International Journal on Digital Libraries [5]. In the future, we plan to continue this series of workshops and Shared Tasks at prominent IR, NLP and Digital Libraries venues.

5. ORGANIZING COMMITTEE

Philipp Mayr is a deputy department head and a team leader at the GESIS – Leibniz-Institute for the Social Sciences department Knowledge Technologies for the Social Sciences (WTS). Philipp Mayr received his PhD in applied informetrics and information retrieval from the Berlin School of Library and Information Science at Humboldt University Berlin in 2009. To date, he has been awarded substantial research funding (PI, Co-PI) from national and European funding agencies. Philipp Mayr has published in top conferences and prestigious journals in the areas informetrics, information retrieval and digital libraries. His research group focuses on methods and techniques for interactive information retrieval. Philipp Mayr was the main organizer of the Combining Bibliometrics and Information Retrieval at ISSI 2013, the BIR workshops at ECIR 2014, 2015, 2016 and 2017 and the BIRNDL workshops at JCDL 2016 and SIGIR 2017.

Kokil Jaidka is a postdoctoral researcher in Computer Science and Chief Technology Officer for the World Wellbeing Project at the University of Pennsylvania. She has been the lead coordinator of all aspects of the CL-SciSumm Shared Task since 2014, and she also co-organized the 1^{st} BIRNDL workshop. She has expertise working on large datasets using machine learning and unsupervised approaches on textual data, and in the specific areas of multi-document summarization and applied linguistics. She is a reviewer for ACL, JCDL, Scientometrics, Applied Linguistics and Aslib journal of Information Processing & Management. Her PhD dissertation involved the development of a literature review

[3]http://clair.eecs.umich.edu/aan/index.php

framework for the summarization of research papers. Currently, she is applying computational methods on social media data for opinion mining, behavioral profiling and modeling health outcomes.

Muthu Kumar Chandrasekaran is broadly interested in natural language processing, machine learning and their applications to information retrieval; specifically, in retrieving and organising information from asynchronous conversation media such as scholarly publications, discussion and debate forums. He has been co-organizing the CL-SciSumm Shared Task series and the BIRNDL workshop series since 2014. He also reviews for ACL, EMNLP, NAACL and JCDL conferences. He believes communication of scholarly research needs to be summarized to avoid redundant or outdated research and ensure faster progress to pressing problems. He is currently doing his Ph.D. research on a similarly motivated problem on Massive Open Online Course (MOOC) discussion forums on recommending salient student discussions for instructors to intervene given their limited bandwidth.

The main organizers will be supported by our previous co-organizers: Guillaume Cabanac, Ingo Frommholz, Min-Yen Kan and Dietmar Wolfram.

6. PROGRAM COMMITTEE

Below, we list the confirmed committee members who have stated their support to review submissions to the workshop at SIGIR 2017. We plan to have three reviews for each BIRNDL submission.

- Akiko Aizawa, National Inst. of Informatics, Japan
- Colin Batchelor, Royal Society of Chemistry, Cambridge, UK
- Joeran Beel, University of Konstanz, Germany
- Marc Bertin, Univ. Claude Bernard Lyon 1, France
- Cornelia Caragea, University of North Texas, USA
- Jason S Chang, National Tsing Hua University, Taiwan
- John Conroy, IDA Center Comput. Sciences, USA
- C Lee Giles, Penn State University, USA
- Bela Gipp, University of Konstanz, Germany
- Nazli Goharian, Georgetown University, USA
- Sujatha Das Gollapalli, Institute for Infocomm Research, A*STAR, Singapore
- Pawan Goyal, Indian Institute of Technology, Kharagpur, India
- Rahul Jha, Microsoft, USA
- Noriko Kando, National Inst. of Informatics, Japan
- Dain Kaplan, Tokyo Institute of Technology, Japan
- Roman Kern, Graz University of Technology, Austria
- Anna Korhonen, University of Cambridge, UK
- Birger Larsen, Aalborg University Copenhagen, Denmark
- John Lawrence, University of Dundee, UK
- Elizabeth Liddy, Syracuse University, USA
- Chin-Yew Lin, Microsoft Research, USA
- Xiaozhong Liu, Indiana Univ., Bloomington, USA
- Kathy McKeown, Columbia University, USA
- Prasenjit Mitra, Penn State University / Qatar Computing Research Institute, USA/Qatar
- Marie-Francine Moens, KU Leuven, Belgium
- Preslav Nakov, Qatar Comp. Research Inst., Qatar
- Doug Oard, Univ. of Maryland, College Park, USA
- Manabu Okumura, Tokyo Inst. of Technology, Japan
- Arzucan Ozgur, Bogazici University, Turkey
- Cecile Paris, CSIRO, Australia
- Vivek Kumar Singh, Banaras Hindu University, India
- Kazunari Sugiyama, National University of Singapore, Singapore
- Simone Teufel, University of Cambridge, UK
- Mike Thelwall, University of Wolverhampton, UK
- Lucy Vanderwende, Microsoft Research, USA
- Vasudeva Varma, International Institute of Information Technology, Hyderabad, India
- Andre Vellino, University of Toronto, Canada
- Anita de Waard, Elsevier Labs, USA
- Alex Wade, Microsoft Research, USA
- Stephen Wan, CSIRO ICT Centre, Australia

7. REFERENCES

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

[2] 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. The computational linguistics summarization pilot task. In *Proceedings of Text Analysis Conference*, Gaithersburg, USA, 2014.

[3] Philipp Mayr, Muthu Kumar Chandrasekaran, and Kokil Jaidka. Report on the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017). *SIGIR Forum*, 51(2):107–113, 2017.

[4] Philipp Mayr, Ingo Frommholz, and Guillaume Cabanac. Report on the 5th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017). *SIGIR Forum*, 51(1):29–35, 2017.

[5] Philipp Mayr, Ingo Frommholz, Guillaume Cabanac, Muthu Kumar Chandrasekaran, Kokil Jaidka, Min-Yen Kan, and Dietmar Wolfram. Introduction to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). *International Journal on Digital Libraries*, 2017.

[6] Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, pages 81–88, 2004.

[7] Simone Teufel. Do "future work" sections have a purpose? citation links and entailment for global scientometric questions. In *Proc. of the 2nd BIRNDL Workshop at SIGIR 2017*, pages 7–13, 2017.

[8] Dietmar Wolfram. Bibliometrics, information retrieval and natural language processing: Natural synergies to support digital library research. In *Proc. of the BIRNDL Workshop 2016*, pages 6–13, 2016.