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

Muthu Kumar Chandrasekaran
School of Computing,
National University of Singapore, Singapore
muthu.chandra@comp.nus.edu.sg

Kokil Jaidka
School of Arts & Sciences,
University of Pennsylvania, USA
jaidka@sas.upenn.edu

Philipp Mayr
GESIS – Leibniz Institute for the Social
Sciences, Germany
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 natural language processing (NLP) techniques could enhance scholarly search, retrieval and user experience 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) [1, 3]. The workshop is intended to stimulate IR, NLP 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 4th edition of the *Computational Linguistics (CL) Scientific Summarization Shared Task*.

1 INTRODUCTION

Over the past several years and at major conferences for information retrieval (IR) research, the BIRNDL workshops [1, 3] and its parent workshops have been establishing themselves as the primary interdisciplinary venue for cross-pollination of natural language processing (NLP), IR and bibliometrics research. The workshop series is motivated by the observation that while the membership in either community shares only a partial overlap the main discourse in both fields consist 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 3rd BIRNDL workshop[1] and 4th CL-SciSumm Shared Task will be a follow-up to the 2nd BIRNDL workshop[2] and 3rd CL-SciSumm Shared Task, co-located with SIGIR 2017, where 7 research papers and 9 system papers were presented[3] [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, 2017 [4] and 2018 at ECIR 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 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

---

[1]http://wing.comp.nus.edu.sg/birndl-sigir2018/
[2]http://wing.comp.nus.edu.sg/birndl-sigir2017/
[3]http://ceur-ws.org/Vol-1888/ and http://ceur-ws.org/Vol-2002/

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

BIRNDL 2018 will have a similar schedule as BIRNDL 2017 at SIGIR. The workshop will start with a keynote titled *"Semi-Automating Biomedical Evidence Synthesis via Machine Learning and Natural Language Processing"* by Prof. Byron Wallace (Northeastern University) followed by regular research paper presentations. The Shared Task overview and selected presentations of the participating teams in the Shared Task will follow. In a poster session 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. The workshop will end with a planning and discussion session to decide on further directions and enhancements to the workshop and the Shared Task.

2.2 The CL-SciSumm Shared Task

The 4th Computational Linguistics (CL) Scientific Summarization Shared Task[4] 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

---

[4]http://wing.comp.nus.edu.sg/cl-scisumm2018/

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)[5]

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.

- 1st 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%).
- 7th Workshop on Bibliometric-enhanced Information Retrieval (BIR2018) at ECIR 2018. The focus of the BIR workshops at ECIR (2014, 2015, 2016, 2017 and 2018) was on research papers in information retrieval, information seeking, science modelling, network analysis, and digital libraries, applying insights from bibliometrics, scientometrics, and informetrics.
- 2nd 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 [2], where the results from 3 system papers were presented.

The following workshops have been organized by other research groups.

- 6th 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.

- 3rd 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. As an output of BIRNDL 2016, we published a special issue on "Bibliometrics, Information Retrieval and Natural Language Processing in Digital Libraries" in the International Journal on Digital Libraries [5]. Another special issue on "Bibliometric-enhanced Information retrieval and Scientometrics" is in preparation for the *Scientometrics* journal. In the future, we plan to continue this series of workshops and Shared Tasks at prominent IR, NLP and Digital Libraries venues.

Since 2016 we maintain the "Bibliometric-enhanced-IR Bibliography" [6] which collects scientific papers which appear in collaboration with the BIR/BIRNDL organizers.

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 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 was the main organizer of the Combining Bibliometrics and Information Retrieval at ISSI 2013, the BIR workshops at ECIR 2014, 2015, 2016, 2017 and 2018 and the BIRNDL workshops at JCDL 2016 and SIGIR 2017. More info: https://philippmayr.github.io/.

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 1st 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 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.

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
- Waleed Ammar, Allen Inst. for Artificial Intelligence, USA
- Iana Atanassova, Universite de Franche-Comte, France
- Colin Batchelor, Royal Society of Chemistry, Cambridge, UK
- Joeran Beel, University of Konstanz, Germany
- Patrice Bellot, Aix-Marseille University, France
- Marc Bertin, Univ. Claude Bernard Lyon 1, France
- Chandra Bhagavatula, Allen Inst. for Artificial Intelligence, USA
- Katarina Boland, GESIS, Germany
- Arman Cohan, Georgetown University, USA
- Anita Dewaard, Elsevier, USA
- Ed A. Fox, Virginia Tech, USA
- Norbert Fuhr, University of Duisburg-Essen, Germany
- Lee Giles, Penn State University, USA
- Pawan Goyal, Indian Institute of Technology, India
- Wolfgang Glaenzel, KU Leuven, Belgium
- Gilles Hubert, University of Toulouse, France
- Dain Kaplan, Tokyo Institute of Technology, Japan
- Roman Kern, Graz University of Technology, Austria
- Atsushi Keyaki, Tokyo Institute of Technology, Tokyo, Japan
- Cyril Labbe, Univ. Grenoble Alpes, France
- Birger Larsen, Aalborg University Copenhagen, Denmark
- John Lawrence, University of Dundee, UK
- Marie-Francine Moens, KU Leuven, Belgium
- Stasa Milojevic, Indiana University, USA
- Preslav Nakov, Qatar Comp. Research Inst., Qatar
- Manabu Okumura, Tokyo Inst. of Technology, Japan
- Byung-won On, Kunsan National University
- Cecile Paris, CSIRO, Australia
- Ameni Sahraoui, GESIS, Germany
- Philipp Schaer, TH Koeln, Germany
- Andrea Scharnhorst, DANS, the Netherlands
- Vivek Kumar Singh, Banaras Hindu University, India
- Kazunari Sugiyama, National Univ. of Singapore, Singapore
- Pradeep Teregowda, IBM, Watson Discovery Services, USA
- Simone Teufel, University of Cambridge, UK
- Mike Thelwall, University of Wolverhampton, UK
- Bart Thijs, KU Leuven, Belgium
- Lucy Vanderwende, Microsoft Research, USA
- Anita de Waard, Elsevier Labs, USA
- Stephen Wan, CSIRO ICT Centre, Australia

REFERENCES

[1] 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

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

[3] 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

[4] Philipp Mayr, Ingo Frommholz, and Guillaume Cabanac. 2017. Report on the 5th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017). *SIGIR Forum* 51, 1 (2017), 29–35. http://sigir.org/wp-content/uploads/2017/06/p029.pdf

[5] Philipp Mayr, Ingo Frommholz, Guillaume Cabanac, Muthu Kumar Chandrasekaran, Kokil Jaidka, Min-Yen Kan, and Dietmar Wolfram. 2017. Introduction to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). *International Journal on Digital Libraries* (2017). https://doi.org/10.1007/s00799-017-0230-x

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

[7] Simone Teufel. 2017. 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.* 7–13. http://ceur-ws.org/Vol-1888/paper1.pdf

[8] 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