# Editorial for the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) at SIGIR 2018

Philipp Mayr[1], Muthu Kumar Chandrasekaran[2], and Kokil Jaidka[3]

[1] GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany,
philipp.mayr@gesis.org
[2] School of Computing,
National University of Singapore, Singapore,
muthu.chandra@comp.nus.edu.sg
[3] School of Arts & Sciences,
University of Pennsylvania, USA,
jaidka@sas.upenn.edu

## 1 Introduction

The goal of the BIRNDL workshop at SIGIR is to engage the information retrieval (IR), natural language processing (NLP), digital libraries, bibliometrics and scientometrics communities to advance the state-of-the-art in scholarly document understanding, analysis and search and retrieval at scale [1]. Scholarly documents are indexed by large, cross-domain digital repositories such as the ACL Anthology, ArXiv, ACM Digital Library, IEEE database, Web of Science, Google Scholar and Semantic Scholar. Currently, digital libraries collect and allow access to papers and their metadata — including 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. Information seeking and sensemaking from the large body of scholarly literature is the key theme of BIRNDL and sets the agenda for tools and approaches to be discussed and evaluated at the workshop.

Papers at the $3^{rd}$ BIRNDL workshop incorporate insights from IR, bibliometrics and NLP to develop new techniques to address the open problems 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 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 are organizing the $4^{th}$ CL-SciSumm Shared Task based on the CL-SciSumm corpus, which comprises over 500 computational linguistics (CL) research papers, interlinked through a citation network.

## 2 Overview of the papers

This year six full papers were submitted to the workshop, three of which were finally accepted as full papers for presentation and inclusion in the proceedings. In addition three poster papers were accepted for inclusion in the proceedings. The workshop featured one keynote talk, one full paper session, one session with presentations of systems participating in the CL-SciSumm Shared Task (see the CL-SciSumm overview paper [2]) and a poster session. The following section briefly describes the keynote and sessions.

### 2.1 Keynote

Byron Wallace gave an inspiring keynote on "Automating Biomedical Evidence Synthesis: Recent Work and Directions Forward" [3]. He talked about recent progress in biomedical evidence synthesis and called on the NLP, IR and machine learning communities to take up the challenges that remain unaddressed in this critical field. He said that the field puts forth technically challenging problems of interest such as, building models with low-supervision, joint inference and extraction over long documents and hybrid crowd-expert annotations, to the aforementioned technical communities.

### 2.2 Research papers

Alzogbi [4] presented a time-aware recommender system (Time-aware Collaborative Topic Regression - T-CTR) that accounts for the concept-drift in user interest by computing user-specific concept drift scores. The paper considered the use case of scientific papers recommendation and conducted experiments on data from citeulike. Results showed the superiority of the time-aware recommendation system T-CTR over the state-of-the-art systems.

Shinoda and Aizawa [5] proposed an unsupervised query-based summarization of scientific papers. Importance scores for words calculated from word embeddings trained on an auxiliary corpus are used to compute sentence vectors. Finally, a random walk is performed on sentences which leverages distributional similarities between query terms and words in the sentence, as well as the similarities between pairs of sentences.

Brochier et al. [6] applied a new document-query methodology to evaluate experts retrieval from a set of queries sampled directly from the experts documents. They provided a formal definition of the expert finding task and worked on a topic-query and a document-query evaluation protocol. They performed a detailed evaluation with three baseline expert recommender algorithms on two AMiner expert data sets.

### 2.3 Poster papers

Scharpf et al. [7] proposed a Wikidata markup to link semantic elements of a mathematical formula in MathML to Wikidata items. They suggested Formula

Concept Discovery as a concept to develop automatic retrieval functions (e.g. formula clustering) on labeled full text corpora. They argued in favor of a larger MathML benchmark for evaluation purposes.

Jia and Saule [8] proposed "Keyphraser" to alleviate the "over-generation error" when extracting key phrases from scientific documents. KeyPhraser is an unsupervised method for identifying document phrases using features such as concordance, popularity, informativeness and position of the first occurrence of the phrase.

Luan et al. [9] presented their SemEval-2018 Task 7 system Semantic Relation Extraction and Classification in Scientific Papers as an invited poster. They were invited due to the work's close relevance to the workshop and the authors' interest.[4]

### 2.4 CL-SciSumm

We hosted the $4^{th}$ Computational Linguistics Scientific Summarization Shared Task, sponsored by Microsoft Research Asia as part of the BIRNDL workshop. The Shared Task is aimed at creation of an open corpus for citation based faceted summarization of scientific documents and evaluation of systems over three sub-tasks to output a summary. This is the first medium-scale shared task on scientific document summarization in the computational linguistics (CL) domain. The task and its corpus have the potential to spur further interest in related problems in scientific discourse mining, such as citation analysis, query-focused question answering and text reuse.

We have been incrementally building our annotated corpora over the four editions of CL-SciSumm. CL-SciSumm'18 systems were provided with 50 document sets for training and evaluation was run over 10 test document sets. Eleven teams registered and submitted their system for evaluation. Results of the evaluation are presented in the CL-SciSumm overview paper [2].

## 3 Outlook and further reading

With this continuing workshop series we have built up a sequence of explorations, visions, results documented in scholarly discourse, and created a sustainable bridge between bibliometrics, IR and NLP. We see the community still growing.

We will continue to organize these kind of workshops at IR, DL, Scientometric, NLP and CL high profile venues. The combination of research paper presentations, and a shared task like CL-SciSumm with system evaluation has proven to be a successful and agile format, so we try to keep this.

In 2015 we published a first special issue on "Combining Bibliometrics and Information Retrieval" in the *Scientometrics* journal [10]. A special issue on "Bibliometrics, Information Retrieval and Natural Language Processing in Digital Libraries" will appear in 2018 in the *International Journal on Digital Libraries*, see

[4] BIRNDL organising committee and [9] thank the BIRNDL reviewers for reviewing this accepted SemEval paper without prejudice.

an overview in [11]. Another special issue on "Bibliometric-enhanced Information Retrieval and Scientometrics" is in preparation for the *Scientometrics* journal (see all accepted papers [5]). Since 2016 we maintain the "Bibliometric-enhanced-IR Bibliography"[6] that collects scientific papers which appear in collaboration with the BIR/BIRNDL organizers. We invite interested researchers to join this project and contribute related publications.

## Acknowledgments

We thank Microsoft Research Asia for their generous support in funding the development, dissemination and organization of the CL-SciSumm dataset and the Shared Task[7]. We are also grateful to the co-organizers of the 1$^{st}$ BIRNDL workshop - Guillaume Cabanac, Ingo Frommholz, Min-Yen Kan and Dietmar Wolfram, for their continued support and involvement. Finally we thank our programme committee members who did an excellent reviewing job. All PC members are documented on the BIRNDL website[8].

## References

1. Mayr, P., Chandrasekaran, M.K., Jaidka, K.: Report on the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017). SIGIR Forum **51**(3) (2017) 107–113
2. Jaidka, K., Yasunaga, M., Chandrasekaran, M.K., Radev, D., Kan, M.Y.: The cl-scisumm shared task 2018: Results and key insights. In: Proc. of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). (2018)
3. Wallace, B.: Automating biomedical evidence synthesis: Recent work and directions forward. In: Proc. of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). (2018)
4. Alzogbi, A.: Time-aware collaborative topic regression: Towards higher relevance in textual item recommendation. In: Proc. of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). (2018)
5. Shinoda, K., Aizawa, A.: Query-focused scientific paper summarization with localized sentence representation. In: Proc. of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). (2018)
6. Brochier, R., Guille, A., Rothan, B., Velcin, J.: Impact of the query set on the evaluation of expert finding systems. In: Proc. of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). (2018)

[5] https://github.com/PhilippMayr/Bibliometric-enhanced-IR_Bibliography/blob/master/bibtex/scientometrics2018.bib

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

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

[8] http://wing.comp.nus.edu.sg/birndl-sigir2018/

7. Scharpf, P., Schubotz, M., Gipp, B.: Representing mathematical formulae in content mathml using wikidata. In: Proc. of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). (2018)
8. Jia, H., Saule, E.: Addressing overgeneration error: An effective and efficient approach to keyphrase extraction from scientific papers. In: Proc. of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). (2018)
9. Luan, Y., Ostendorf, M., Hajishirzi, H.: The UWNLP system at SemEval-2018 Task 7: Neural Relation Extraction Model with Selectively Incorporated Concept Embeddings. In: Proceedings of The 12th International Workshop on Semantic Evaluation, ACL (2018) 788–792
10. Mayr, P., Scharnhorst, A.: Scientometrics and information retrieval: weak-links revitalized. Scientometrics **102**(3) (2015) 2193–2199
11. Mayr, P., Frommholz, I., Cabanac, G., Chandrasekaran, M.K., Jaidka, K., Kan, M.Y., Wolfram, D.: Introduction to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). International Journal on Digital Libraries (2017)