Report on the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017)

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

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

Kokil Jaidka
University of Pennsylvania, USA
jaidka@sas.upenn.edu

Abstract

The large scale of scholarly publications poses a challenge for scholars in information seeking and sense-making. Bibliometrics, information retrieval (IR), text mining, and NLP techniques could help in these activities, but are not yet widely implemented in digital libraries. The 2^{nd} joint BIRNDL workshop was held at the 40th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017) in Tokyo, Japan. BIRNDL 2017 intended to stimulate IR researchers and digital library professionals to elaborate on new approaches in natural language processing, information retrieval, scientometric, and recommendation techniques that can advance the state-of-the-art in scholarly document understanding, analysis, and retrieval at scale. The workshop incorporated three paper sessions and the 3^{rd} edition of the CL-SciSumm Shared Task.

1 Introduction

Over the past several years, the BIRNDL workshop and its parent workshops are establishing themselves as the primary interdisciplinary venue for the cross-pollination of bibliometrics and information retrieval (IR) [16]. Our motivation as organizers of the workshop started from the observation that both communities share only a partial overlap; yet, the main discourse in both fields consists of different approaches to solve similar problems. We believe that a knowledge transfer would be profitable for both sides. A good overview of the symbiotic relationship that exists among bibliometrics, IR and natural language processing (NLP) has been presented last year by Wolfram [21]. A report of the first BIRNDL workshop has been published in the SIGIR Forum [4].

The goal of the BIRNDL workshop at SIGIR is to engage the IR community about the open problems in academic search. Academic search refers to the large, cross-domain digital

repositories which index research papers, such as the ACL Anthology, ArXiv, ACM Digital Library, IEEE database, Web of Science and Google 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. Finding relevant 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 2^{nd} 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 3^{rd} 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 14 papers were submitted to the workshop, 5 of which were finally accepted as full papers and 2 were accepted as short papers for presentation and inclusion in the proceedings[1]. In addition 3 poster papers were accepted. The workshop featured one keynote talk, two paper sessions, one session with presentations of systems participating in the CL-SciSumm Shared Task and a poster session. The following section briefly describes the keynote and sessions.

2.1 Keynote

The invited paper "Do "Future Work" sections have a purpose? Citation links and entailment for global scientometric questions" [20] by Simone Teufel (University of Cambridge, UK) gives new perspectives basing on NLP techniques on the "Future Works" sections in scientific papers. The author raises questions like: Where is the research of a field going? Where are the currently most challenging research issues? Where are the future game-changers? The author ends with a nexus to scientometric applications like citation function classification. Simone Teufel argues that scientometric research could and should be connected and complemented more with computational linguistics.

2.2 Session 1

The paper "Can we do better than Co-Citations? - Bringing Citation Proximity Analysis from idea to practice in research article recommendation" by Knoth and Khadka [12] describes a practical approach, namely research article recommendation, that builds on Citation Proximity Analysis (CPA) (a Co-Citation approach defining a high co-citedness index as a high relatedness). The authors built a CPA-based recommender system from a large

[1] http://ceur-ws.org/Vol-1888/

corpus of full-texts articles from the CORE text corpus and conducted a user survey to perform an initial evaluation. Two of the three proximity functions used within CPA outperform co-citations on their evaluation dataset.

The paper "MultiScien: a Bi-Lingual Natural Language Processing System for Mining and Enrichment of Scientific Collections" by Saggion, Ronzano, Accuosto and Ferres describes MultiScien – a system for deep analysis and annotation of research papers, and introduces the SEPLN anthology, an annotated bilingual corpus of SEPLN publications [19]. The authors address the specific challenges involved in mining bi-lingual text from the formatting layout particular to SEPLN publications.

The paper "Identifying Problems and Solutions in Scientific Text" by Heffernan and Teufel [9] proposes an automatic classier that makes a binary decision about "problemhood" and "solutionhood" of a given phrase from a scientific paper. They treated the problem as a supervised machine learning problem and evaluated their approach on the basis of an own corpus (a subset of the latest version of the ACL anthology) consisting of 2,000 positive and negative examples of problems and solutions. According to their evaluation, part of speech (POS) tags and document and word embeddings are the best performing features.

2.3 Session 2

Caglier *et* al. [5] address the problem of mining collaborations patterns to measure their impact on research areas or *topics*. In their paper "Identifying Collaborations among Researchers: a pattern-based approach" they draw upon established data mining algorithm, frequent-itemset mining to discover author-topic patterns that frequently co-occur.

The paper "Automatic Generation of Review Matrices as Multi-document Summarization of Scientific Papers" by Hashimoto, Shinoda, Yokono and Aizawa [8] describes a summarization system to generate a synthesis matrix from an overview of closely-related papers. They formulate the problem as a query-focused summarization problem and use lexical ranking methods to order and select the most appropriate sentences which describe an aspect of a cited paper.

The paper by Bar-Ilan "Bibliometrics of Information Retrieval – A Tale of Three Databases" [2] studies coverage issues of the three bibliographic databases Web of Science (WoS), Scopus and the ACM Digital Library. The paper shows a rather small overlap between the results retrieved by the databases. Only 12% of the retrieved documents were covered by all three databases.

The paper "Analysis of Footnote Chasing and Citation Searching in an Academic Search Engine" by Kacem and Mayr [10] analyzes the user behaviour towards Marcia Bates' search stratagems 'footnote chasing' and 'citation search' in a large logfile of the academic search engine in the social sciences, called sowiport. They showed that the appearance of 'footnote chasing' and 'citation search' in real interactive retrieval sessions lead to an improvement of the precision in terms of positive signals like (downloading, exporting or sharing) after using these stratagems.

2.4 Session 3: CL-SciSumm

As a part of the workshop, we conducted the 3^{rd} Computational Linguistics Scientific Summarization Shared Task, sponsored by Microsoft Research Asia. This is the 3^{rd} edition first

medium-scale shared task on scientific document summarization in the computational linguistics (CL) domain. It is based on an annotated corpus of 40 topics, each comprising 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 have been identified. Participants were required to solve three sub-tasks in automatic research paper summarization on a text corpus. Ten teams participated and completed 58 submissions to the tasks, which employed a variety of lexical and graph-based features in unsupervised and supervised approaches. Six of these teams had previously participated in the 2^{nd} CL-SciSumm Shared Task at BIRNDL 2016 [4]. 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 list here the final system reports of the participant systems[2]: [15] [18] [14] [11] [17] [6] [13] [22]. Readers are directed to refer to the system reports for their methodologies and results.

2.5 Poster session

Hamborg *et* al. [7] propose a method for automatically generating patent abstracts and time-stamping them in their bid to stop patent trolls from filing obvious patents.

Bertin and Atanassova [3] introduce an approach to explore the multidimensional nature of the elements composing the contexts of citations in different sections of research papers, based on unsupervised clustering of a random sample of citing sentences from seven peer-reviewed open-access academic journals.

Alam *et* al. [1] describe a simple cosine-similarity based proof-of-concept system to evaluate textual similarity between reference spans and citing texts of pairs of papers. This paper was invited for a poster presentation at the workshop to encourage industry participation in digital library and bibliometrics research since the industry runs some of the largest and widely used bibliometrics and digital library systems (e.g., Google Scholar).

3 Outlook

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.

This year, the authors of accepted papers at the 2nd BIRNDL workshop were invited to submit extended versions to a Special Issue on "Bibliometric-enhanced IR" of the *Scientometrics*[3] journal to be published in 2018. After the first BIRNDL workshop at JCDL 2016 we started a Special Issue in the *International Journal on Digital Libraries*[4]. The production of this issue is currently in process. All accepted and published papers are documented in a bibtex file[5].

[2]The CL-SciSumm Shared Task system reports will appear in volume 2 of the BIRNDL workshop proceedings published by CEUR and are documented on the BIRNDL website.

[3]`http://www.springer.com/journal/11192`

[4]`https://link.springer.com/journal/799`

[5]https://github.com/PhilippMayr/Bibliometric-enhanced-IR_Bibliography/blob/master/bibtex/ijdl2017.bib

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.

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[6]. 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[7].

References

[1] H. Alam, A. Kumar, T. Werner, and M. Vyas. Are Cited References Meaningful? Measuring Semantic Relatedness in Citation Analysis. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 113–118, Tokyo, Japan, 2017. CEUR-WS.org.

[2] J. Bar-Ilan. Bibliometrics of "Information Retrieval" – A Tale of Three Databases. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 83–90, Tokyo, Japan, 2017. CEUR-WS.org.

[3] M. Bertin and I. Atanassova. K-means and Hierarchical Clustering Method to Improve our Understanding of Citation Contexts. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 107–112, Tokyo, Japan, 2017. CEUR-WS.org.

[4] G. Cabanac, M. K. Chandrasekaran, I. Frommholz, K. Jaidka, M.-Y. Kan, P. Mayr, and D. 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.

[5] L. Cagliero, P. Garza, M. R. Kavoosifar, and E. Baralis. Identifying collaborations among researchers: a pattern-based approach. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 56–68, Tokyo, Japan, 2017. CEUR-WS.org.

[6] T. Felber and R. Kern. TU-GRAZ@CL-SciSumm17 Query Generation Strategies for CL-SciSumm 2017 Shared Task. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2*, Tokyo, Japan, 2017. CEUR-WS.org.

[6] `http://wing.comp.nus.edu.sg/cl-scisumm2017/`

[7] `http://wing.comp.nus.edu.sg/birndl-sigir2017/`

[7] F. Hamborg, M. Elmaghraby, C. Breitinger, and B. Gipp. Automated Generation of Timestamped Patent Abstracts at Scale to Outsmart Patent-Trolls. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 101–106, Tokyo, Japan, 2017. CEUR-WS.org.

[8] H. Hashimoto, K. Shinoda, H. Yokono, and A. Aizawa. Automatic Generation of Review Matrices as Multi-document Summarization of Scientific Papers. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 69–82, Tokyo, Japan, 2017. CEUR-WS.org.

[9] K. Heffernan and S. Teufel. Identifying Problems and Solutions in Scientific Text. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 41–55, Tokyo, Japan, 2017. CEUR-WS.org.

[10] A. Kacem and P. Mayr. Analysis of Footnote Chasing and Citation Searching in an Academic Search Engine. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 91–100, Tokyo, Japan, 2017. CEUR-WS.org.

[11] S. Karimi, R. Verma, L. Moraes, and A. Das. U.Houston@CL-SciSumm17: Correspondence Learning and Textual Entailment. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2*, Tokyo, Japan, 2017. CEUR-WS.org.

[12] P. Knoth and A. Khadka. Can we do better than Co-Citations? - Bringing Citation Proximity Analysis from idea to practice in research article recommendation. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 14–25, Tokyo, Japan, 2017. CEUR-WS.org.

[13] A. Lauscher, G. Glavaš, and K. Eckert. University of Mannheim @ CLSciSumm-17: Citation-Based Summarization of Scientific Articles Using Semantic Textual Similarity. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2*, Tokyo, Japan, 2017. CEUR-WS.org.

[14] L. Li, L. Mao, Y. Zhang, J. Chi, and M. Chen. CIST@CL-SciSumm17: Multiple Features Based Citation Linkage, Classifiation and Summarization. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2*, Tokyo, Japan, 2017. CEUR-WS.org.

[15] S. Ma, J. Wang, J. Xu, and C. Zhang. NJUST@CL-SciSumm17. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2*, Tokyo, Japan, 2017. CEUR-WS.org.

[16] P. Mayr and A. Scharnhorst. Scientometrics and Information Retrieval - weak-links revitalized. *Scientometrics*, 102(3):2193–2199, 2015.

[17] A. Prasad. WING-NUS@CL-SciSumm17: Learning from Syntactic and Semantic Similarity for Citation Contextualization. In *Proc. of the 2nd Joint Workshop on*

Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2, Tokyo, Japan, 2017. CEUR-WS.org.

[18] H. Saggion, A. Aburaed, and L. Chiruzzo. LaSTUS/TALN @ CLSciSumm-17:Cross-document Sentence Matching and Scientific Text Summarization Systems. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2*, Tokyo, Japan, 2017. CEUR-WS.org.

[19] H. Saggion, F. Ronzano, P. Accuosto, and D. Ferrés. MultiScien: a Bi-Lingual Natural Language Processing System for Mining and Enrichment of Scientific Collections. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 26–40, Tokyo, Japan, 2017. CEUR-WS.org.

[20] S. Teufel. Do "Future Work" sections have a real purpose? Citation links and entailment for global scientometric questions. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 7–13, Tokyo, Japan, 2017. CEUR-WS.org.

[21] D. 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.

[22] D. Zhang and S. Li. PKU @ CLSciSumm-17: Citation Contextualization. In *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Vol. 2*, Tokyo, Japan, 2017. CEUR-WS.org.