Report on the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries at SIGIR 2019

Muthu Kumar Chandrasekaran
Independent Researcher, CA, USA
cmkumar087@gmail.com

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

Abstract

The 4^{th} joint BIRNDL workshop was held at the 42nd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019) in Paris, France. BIRNDL 2019 intended to stimulate IR researchers and digital library professionals to elaborate on new approaches in natural language processing, information retrieval, scientometrics, and recommendation techniques that can advance the state-of-the-art in scholarly document understanding, analysis, and retrieval at scale. The workshop incorporated different paper sessions and the 5^{th} edition of the CL-SciSumm Shared Task.

1 Introduction

The goal of the BIRNDL workshop at SIGIR 2019 was to engage the IR community in the open problems in Big Science. Big Science refers to the large, cross-domain digital repositories which index research papers, such as the ACL Anthology, ArXiv, ACM Digital Library, Semantic Scholar, PubMed, IEEE database, Web of Science, Google Scholar etc. 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 growth in scholarly publishing 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 methods and approaches to be discussed and evaluated at BIRNDL. 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 organised the 4^{th} Joint Workshop on Bibliometric-enhanced Information Retrieval and

Natural Language Processing for Digital Libraries (BIRNDL)[1] and the 5^{th} CL-SciSumm Shared Task[2] co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019).

Over the past several years and at premier conferences, BIRNDL [1, 2, 3], together with its parent workshops, has established itself as the primary interdisciplinary venue for cross-pollination of NLP, IR and DL research. 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. For 2019, we invited 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 included:

- Infrastructure for scientific mining and IR,
- Semantic and network-based indexing, search and navigation in structured text,
- Discourse modeling and argument mining,
- Summarization and question-answering for scholarly DLs,
- Bibliometrics and citation analysis,
- 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.

2 Overview of the papers

This year, 11 full papers were submitted to the workshop, 5 of which were accepted as full papers for presentation and inclusion in the proceedings [4]. 13 short papers were submitted to the workshop, 4 of which were accepted as short papers for presentation. In addition, 5 poster papers for poster presentation were also accepted.

The CL-SciSumm Shared Task had 17 registrations of which 9 teams submitted their systems for evaluation. The results and analysis of the evaluation along with the system summaries are presented in the overview paper titled: "Overview and Results: CL-SciSumm Shared Task 2019" [5]. The system description papers are included in the proceedings.

The workshop featured two keynote talks, one invited talk and various full and short paper sessions. The following section briefly summarizes the keynotes, invited talk and full research papers. The short papers, posters and CL-SciSumm papers are just listed.

2.1 Keynotes

- Alex D. Wade, Ivana Williams: *Personalized Feed/Query-formulation, Predictive Impact, and Ranking* [6]. In his keynote, Alex Wade introduced the Meta discovery system

[1] `http://wing.comp.nus.edu.sg/~birndl2019/`
[2] `http://wing.comp.nus.edu.sg/~cl-scisumm2019/`

(https://www.meta.org/) which is a biomedical paper discovery and current awareness service developed by the Chan Zuckerberg Initiative. The system implements state-of-the-art NLP features and involves a knowledge graph derived from biomedical literature harvested from Pubmed and bioRxiv.

- Bonnie Webber: *Discourse Processing for Text Analysis: Recent successes, current challenges* [7]. Bonnie Webber reviewed 10 years of computational discourse processing literature. She summarizes changing assumptions about discourse structure, recent work on lexico-syntactic grounding of low-level discourse structure and points out some remaining challenges.

2.2 Invited Paper

- *TalkSumm: A Dataset and Scalable Annotation Method for Scientific Paper Summarization Based on Conference Talks* [8] which was a short paper accepted at ACL 2019 was presented before ACL at BIRNDL relevance to CL-SciSumm Shared Task. David Konopnicki, heading IBM Research Israel group on scholarly document research spoke about TalkSumm which introduces a scalable method to annotate data for scientific document summarisation from transcripts of talks by authors of the a scientific paper. Further, he also presented a live demonstration of IBM's section-wise summarisation system.
  This paper appeared in *Proceedings of ACL 2019* and was not included in the BIRNDL proceedings.

2.3 Research papers

2.3.1 Full papers

- Katarina Boland, Frank Krüger: *Distant supervision for silver label generation of software mentions in social scientific publications* [9]. The authors investigate in a case study the use of weakly supervised approaches with distant supervision to create silver labels to train supervised software mention extraction methods using transfer learning. They worked with a small sample of social sciences full text papers from the Open Access publisher PLoS.
- Na Pang, Li Qian, Weimin Lyu, Jin-Dong Yang: *Transfer Learning for Scientific Data Chain Extraction in Small Chemical Corpus with joint BERT-CRF Model* [10] proposes an annotated corpus for entity recognition in Chemistry. They propose a transfer learning sequence tagging model, BERT-CRF to extract entities. We believe the corpus will push the state of the art in this important research area.
- Christin Katharina Kreutz, Premtim Sahitaj, Ralf Schenkel: *Revaluating Semantometrics from Computer Science Publications* [11]. The authors try to learn patterns of features extracted from publication distances in their citation networks aiming at distinguishing between seminal and survey papers in the area of computer science. They used the SeminalSurveyDBLP dataset for their evaluations.
- Suzan Verberne, Ioannis Chios, Jian Wang: *Extracting and matching patent in-text references to scientific publications* [12] propose a pipeline to extract citations to scientific articles found in patents. They propose a CRF based pipeline to extract references in

the article rather than citations on the front page of the patent which they show are easier to extract. They show state-of-the-art performance.

- Michael Soprano, Kevin Roitero, Stefano Mizzaro: *HITS Hits Readersourcing: Validating Peer Review Alternatives Using Network Analysis* [13]. The authors propose a stochastic validation of the Readersourcing model as an alternative of the peer review system, and employ network analysis techniques to study the bias of the model. They intend to study the interactions between the skill of a reader and the quality of a paper, where such quantities are computed by Readersourcing models.

2.3.2 Short papers

- Jason Portenoy, Jevin D. West: *Supervised Learning for Automated Literature Review* [14]
- Arlene J Casey, Bonnie Webber, Dorota Glowacka: *Can Models of Author Intention Support Quality Assessment of Content?* [15]
- Andres Carvallo, Denis Parra: *Comparing Word Embeddings for Document Screening based on Active Learning* [16]
- Philipp Scharpf, Moritz Schubotz, Howard S. Cohl, Bela Gipp: *Towards Formula Concept Discovery and Recognition* [17]

2.3.3 Poster papers

- Barbara Plank, Reinard van Dalen: *CiteTracked: A Longitudinal Dataset of Peer Reviews and Citations* [18]
- Chifumi Nishioka, Michael Färber: *Evaluating the Availability of Open Citation Data* [19]
- André Greiner-Petter, Terry Ruas, Moritz Schubotz, Akiko Aizawa, William Grosky, Bela Gipp: *Why Machines Cannot Learn Mathematics, Yet* [20]
- Marc Bertin, Pierre Jonin, Frédéric Armetta, Iana Atanassova: *Identifying the conceptual space of citation contexts using coreferences* [21]
- Rajesh Piryani, Wolfgang Otto, Philipp Mayr, Vivek Kumar Singh: *Analysing author name mentions in citation contexts of highly cited publications* [22]

2.4 CL-SciSumm

CL-SciSumm tasks systems to summarise a reference paper (RP) using contents from the RP and citing sentences from up to 10 papers that cite RP. CL-SciSumm Shared Task consists of 3 sub tasks: Task 1a asks systems to infer cited sentences in the RP given a set of citing sentences from the cited paper as input. Task 1b asks systems to assign a discourse type (1 out of 4) to the cited sentences. Task 2 asks systems to generate a summary of RP. The summary is evaluated against 3 reference summaries: the abstract of the RP, a human written summary and a community summary made from the citing sentences of the citing papers. The annotated dataset used for this shared task and the scripts used for evaluation can be accessed and used by the community at: `https://github.com/WING-NUS/scisumm-corpus`. Nine teams submitted their system for evaluation. They are as follows. *NaCTeM-UoM @*

CL-SciSumm-19 [23] presented their results at BIRNDL since their system was the best on Task 1a and Task 2 when evaluated on abstracts. However, *CIST@CLSciSumm-19: Automatic Scientific Paper Summarization With Citances and Facets* [24] had the best results on Task 1b. They and *Poli2Sum@CL-SciSumm 2019: identify, classify, and summarize cited text spans by means of ensembles of supervised models* [25] had the best results on Task 2 when evaluated against community and human summaries. More detailed results and analysis of CL-SciSumm 2019 are written in our overview paper [26] which we encourage you to read. CL-SciSumm 2020 will feature a new summarisation task and will be proposed at BIRNDL @ SIGIR 2020.

- Chrysoula Zerva, Minh-Quoc Nghiem, Nhung Nguyen and Sophia Ananiadou: *NaCTeM-UoM @ CL-SciSumm-19* [23]
- Shutian Ma, Heng Zhang, Tianxiang Xu, Jin Xu, Shaohu Hu and Chengzhi Zhang: *IR&TM-NJUST @ CLSciSumm-19* [27]
- Lei Li, Yingqi Zhu, Wei Liu, Zuying Huang, Yang Xie, Yinan Liu and Xingyuan Li: *CIST@CLSciSumm-19: Automatic Scientific Paper Summarization With Citances and Facets* [24]
- Yoann Pitarch, Karen Pinel-Sauvagnat, Gilles Hubert, Guillaume Cabanac and Ophélie Fraisier-Vannier: *IRIT-IRIS @ CL-SciSumm 2019: Matching Citances with their Intended Reference Text Spans from the Scientic Literature* [28]
- Bakhtiyar Syed, Vijayasaradhi Indurthi, Balaji Vasan Srinivasan and Vasudeva Varma: *Helium @ CL-SciSumm-19: Transfer learning for effective scientific research comprehension* [29]
- Ahmed Ghassan Tawfiq Abura'Ed, Àlex Bravo, Luis Chiruzzo and Horacio Saggion: *LaSTUS/TALN+INCO @ CL-SciSumm 2019* [30]
- Moreno La Quatra, Luca Cagliero and Elena Baralis: *Poli2Sum@CL-SciSumm 2019: identify, classify, and summarize cited text spans by means of ensembles of supervised models* [25]
- Hyonil Kim and Shiyan Ou: *NJU@CL-SciSumm-19: Ranking-based Identification of Cited Text with Deep Learning* [31]
- Aris Fergadis, Dimitris Pappas and Haris Papageorgiou *ATHENA@CL-SciSumm 2019: Siamese recurrent bi-directional neural network for identifying cited text spans* [32]

3 Outlook and further reading

We will continue to organize the BIRNDL workshops at high profile IR, DL, Scientometric, NLP and CL venues. For 2020, we plan an additional workshop alongside an ACL event. 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 will continue with this format.

In line with the BIRNDL 2019 workshop, we invited workshop authors and others to submit extended paper to two independent open calls of special issues.

1. Special issue on "Mining Knowledge from Scientific Data"[3] in the journal Expert Systems (Wiley). Guest editors are: Tanmoy Chakraborty, Sumit Bhatia and Cornelia Caragea. Deadline: September 30, 2019
2. Special issue on "Bibliometrics and Information Retrieval"[4] in the journal Scientometrics (Springer). Guest editors are: Guillaume Cabanac, Philipp Mayr and Ingo Frommholz. Deadline: September 30, 2019

Further reading: A recent research topic on "Mining Scientific Papers: NLP-enhanced Bibliometrics" appeared in 2019 in *Frontiers in Research Metrics and Analytics*, see an overview in [33]. A special issue on "Bibliometrics, Information Retrieval and Natural Language Processing in Digital Libraries" appeared in 2018 in the *International Journal on Digital Libraries*, see an overview in [34]. This was a large issue with 13 papers with papers from BIRNDL 2016 as well as new submissions. So, we followed up with another special issue on "Bibliometric-enhanced Information Retrieval and Scientometrics" appeared in *Scientometrics* journal with extended papers from later editions BIR and BIRNDL and new submissions (see an overview in [35]).

Since 2016 we maintain the "Bibliometric-enhanced-IR Bibliography"[5] 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 immensely thank Prof. Dragomir Radev and Michihiro Yasunaga from Yale University for sharing the SciSummNet dataset for CL-SciSumm 2019 and co-organising this time. We thank sponsors for sponsoring keynote speakers. CZI sponsored Alex Wade while SRI International partially sponsored Bonnie Webber to attend and deliver her keynote. The 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. Philipp Mayr acknowledges the enabling support provided by the Indo-German Joint Research Project titled 'Design of a Sciento-text Computational Framework for Retrieval and Contextual Recommendations of High-Quality Scholarly Articles' (Grant No. DST/INT/FRG/DAAD/P-28/2017) for this work.

We are also grateful to the co-organizers of the past BIRNDL workshops - Guillaume Cabanac, Ingo Frommholz, Kokil Jaidka and Dietmar Wolfram, for their continued support and involvement.

References

[1] Mayr, P., Chandrasekaran, M.K., Jaidka, K.: 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

[3] http://lcs2.iiitd.edu.in/speIssue_Exsy.html

[4] https://sites.google.com/view/scientometrics-si2019-bir

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

[2] Mayr, P., Chandrasekaran, M.K., Jaidka, K.: 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

[3] Cabanac, G., Chandrasekaran, M.K., Frommholz, I., Jaidka, K., Kan, M.Y., Mayr, P., Wolfram, D.: 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

[4] Chandrasekaran, M.K., Mayr, P., eds.: Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, July 25, 2019. Volume 2414 of CEUR Workshop Proceedings., CEUR-WS.org (2019)

[5] Chandrasekaran, M.K., Yasunaga, M., Radev, D.R., Freitag, D., Kan, M.: Overview and results: Cl-scisumm shared task 2019. [4] 153–166

[6] Wade, A.D., Williams, I.: Personalized feed/query-formulation, predictive impact, and ranking. [4] 6–7

[7] Webber, B.: Discourse processing for text analysis: Recent successes, current challenges. [4] 8–14

[8] Lev, G., Shmueli-Scheuer, M., Herzig, J., Jerbi, A., Konopnicki, D.: TalkSumm: A Dataset and Scalable Annotation Method for Scientific Paper Summarization Based on Conference Talks. In: arXiv:1906.01351

[9] Boland, K., Krüger, F.: Distant supervision for silver label generation of software mentions in social scientific publications. [4] 15–27

[10] Pang, N., Qian, L., Lyu, W., Yang, J.: Transfer learning for scientific data chain extraction in small chemical corpus with joint BERT-CRF model. [4] 28–41

[11] Kreutz, C.K., Sahitaj, P., Schenkel, R.: Revaluating semantometrics from computer science publications. [4] 42–55

[12] Verberne, S., Chios, I., Wang, J.: Extracting and matching patent in-text references to scientific publications. [4] 56–69

[13] Soprano, M., Roitero, K., Mizzaro, S.: HITS hits readersourcing: Validating peer review alternatives using network analysis. [4] 70–82

[14] Portenoy, J., West, J.D.: Supervised learning for automated literature review. [4] 83–91

[15] Casey, A.J., Webber, B., Glowacka, D.: Can models of author intention support quality assessment of content? [4] 92–99

[16] Carvallo, A., Parra, D.: Comparing word embeddings for document screening based on active learning. [4] 100–107

[17] Scharpf, P., Schubotz, M., Cohl, H.S., Gipp, B.: Towards formula concept discovery and recognition. [4] 108–115

[18] Plank, B., van Dalen, R.: Citetracked: A longitudinal dataset of peer reviews and citations. [4] 116–122

[19] Nishioka, C., Färber, M.: Evaluating the availability of open citation data. [4] 123–129

[20] Greiner-Petter, A., Ruas, T., Schubotz, M., Aizawa, A., Grosky, W.I., Gipp, B.: Why machines cannot learn mathematics, yet. [4] 130–137

[21] Bertin, M., Jonin, P., Armetta, F., Atanassova, I.: Identifying the conceptual space of citation contexts using coreferences. [4] 138–144

[22] Piryani, R., Otto, W., Mayr, P., Singh, V.K.: Analysing author name mentions in citation contexts of highly cited publications. [4] 145–152

[23] Zerva, C., Nghiem, M., Nguyen, N.T.H., Ananiadou, S.: Nactem-uom @ cl-scisumm 2019. [4] 167–180

[24] Li, L., Zhu, Y., Xie, Y., Huang, Z., Liu, W., Li, X., Liu, Y.: Cist@clscisumm-19: Automatic scientific paper summarization with citances and facets. [4] 196–207

[25] Quatra, M.L., Cagliero, L., Baralis, E.: Poli2sum@cl-scisumm-19: Identify, classify, and summarize cited text spans by means of ensembles of supervised models. [4] 233–246

[26] Chandrasekaran, M.K., Yasunaga, M., Radev, D., Freitag, D., Kan, M.Y.: Overview and results: Cl-scisumm shared task 2019. arXiv preprint arXiv:1907.09854 (2019)

[27] Ma, S., Zhang, H., Xu, T., Xu, J., Hu, S., Zhang, C.: Ir&tm-njust @ clscisumm-19. [4] 181–195

[28] Pitarch, Y., Pinel-Sauvagnat, K., Hubert, G., Cabanac, G., Fraisier-Vannier, O.: IRIT-IRIS at cl-scisumm 2019: Matching citances with their intended reference text spans from the scientific literature. [4] 208–213

[29] Syed, B., Indurthi, V., Srinivasan, B.V., Varma, V.: Helium @ cl-scisumm-19 : Transfer learning for effective scientific research comprehension. [4] 214–223

[30] Chiruzzo, L., AbuRa'ed, A., Bravo, À., Saggion, H.: Lastus-taln+inco @ cl-scisumm 2019. [4] 224–232

[31] Kim, H., Ou, S.: Nju@cl-scisumm-19. [4] 247–255

[32] Fergadis, A., Pappas, D., Papageorgiou, H.: Athena@cl-scisumm 2019: Siamese recurrent bi-directional neural network for identifying cited text spans. [4] 256–262

[33] Atanassova, I., Bertin, M., Mayr, P.: Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. Frontiers in Research Metrics and Analytics **4**(2) (2019)

[34] 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 **19**(2-3) (2018) 107–111

[35] Cabanac, G., Frommholz, I., Mayr, P.: Bibliometric-enhanced information retrieval: preface. Scientometrics **116**(2) (2018) 1225–1227