Editorial for the Second Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics (CLBib2017)

Iana Atanassova
Centre Tesnière - CRIT, University of Bourgogne Franche-Comté, Besançon, France
iana.atanassova@univ-fcomte.fr
Marc Bertin
ELICO Laboratory, University of Lyon, Lyon, France
marc.bertin@univ-lyon1.fr
Philipp Mayr
GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany
Philipp.Mayr@gesis.org

1 Introduction

The Open Access movement in scientific publishing and search engines like Google Scholar have made scientific articles more broadly accessible. During the last decade, the availability of scientific papers in full text has become more and more widespread thanks to the growing number of publications on online platforms such as ArXiv, CiteSeer and Public Library of Science (PLOS). In this context, new needs arise around the processing and efficient exploitation of scientific corpora.

Scientific papers are highly structured texts and display specific properties related to their references but also argumentative and rhetorical structure. Recent research in this field has concentrated on the construction of ontologies for citations and scientific articles (e.g. FaBiO and CiTO [8]) and studies of the distribution of references (see [2]). However, up to now full-text mining efforts are rarely used to provide data for bibliometric analyses. While bibliometrics traditionally relies on the analysis of metadata of scientific papers (see e.g. a recent special issue on "Combining Bibliometrics and Information Retrieval", Mayr & Scharnhorst [6]), we will explore the ways full-text processing of scientific papers and linguistic analyses can play.

The CLBib workshop series provides a forum to discuss novel approaches and insights into scientific writing that can bring new perspectives to understand both the nature of citations and the nature of scientific articles. The possibility to enrich metadata by the full-text processing of papers offers new fields of application to bibliometrics studies.

2 Scope and Motivation

The CLBib workshops aim to bring together researchers in bibliometrics and computational linguistics in order 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 Natural Language Processing. Working with full text allows us to go beyond metadata used in bibliometrics. Full text offers a new field of investigation, where the major problems arise around the organization and structure of text, the extraction of information and its representation on the level of metadata. Furthermore, the study of contexts around in-text citations offers new perspectives related to the semantic dimension of citations. The analyses of citation contexts and the semantic

In: Iana Atanassova, Marc Bertin and Philipp Mayr (eds.): Proceedings of the CLBib2017 Workshop, Wuhan, China, 17-oct-2017, published at http://ceur-ws.org

categorization of publications will allow us to rethink co-citation networks, bibliographic coupling and other bibliometric techniques.

The first edition of this workshop[1], co-located with the International Society of Scientometrics and Informetrics Conference (ISSI) in 2015, attracted more than 70 participants and six full paper contributions, showing a large interest in these topics in the community. From a technical point of view, during the first edition of the workshop, the efforts to provide articles in machine-readable formats and the rise of Open Access publishing have resulted in a number of standardized formats for scientific papers, full-text datasets for research experiments and corpora and focus on number of open source tools for versatile text processing.

The goal of this second edition of the CLBib workshop, co-located with the ISSI conference 2017, is to continue to encourage the collaboration between these two domains and to answer questions like: How can we enhance author network analysis and Bibliometrics using data obtained by text analytics? What insights can NLP provide on the structure of scientific writing, on citation networks, and on in-text citation analysis? Natural Language Processing and Bibliometrics meet again in this second workshop in a context where Open Access is at the heart of exchanges between scientists and publishers and raises many economic and ethical issues, but also new research problems through the access to articles in full text. Indeed, the possibility of enriching metadata currently used in bibliometrics with information from the text is an essential step towards building the tools of tomorrow.

As the CLBib 2017 workshop was held in China, at Wuhan University, the discussions raised important questions not only around the processing of scientific papers but also on the need to take into account the multilingual aspect of the scientific production. Even if today English is essential on the international stage, national level publications can also be rich in information and relevant for bibliometric studies. The linguistic aspect, which is more and more present at the ISSI conference, must be taken into consideration and highlights the importance of this workshop series and the growing interest in the community of bibliometricians but also in other communities for Natural Language Processing.

3 Overview of the papers

The call for papers[2] attracted several submissions, of which 50% were accepted for publication. The workshop featured an introduction and four paper presentations. All papers are included in the current proceedings. We shortly summarize each workshop paper below. The publications selected for this workshop CLBib2017 concern both, theory and application aspects of bibliometric.

For this second edition of the CLBib workshop, the authors used methods that come from various fields, such as Information Retrieval with traditional measures (e.g. tf-idf, vector models, ...), network analysis and visualisation with Ucinet, Netdraw, CiteSpace (see [3]), and also NLP with word embedding techniques and co-words analysis.

The paper "Understanding the Changing Roles of Scientific Publications via Citation Embeddings" by **Jiangen He and Chaomei Chen** [5] describes an approach which helps to understand the changing and complex role of a publication characterized by its citation contexts. The authors propose a temporal representation of in-text citations of publications as a sequence of vectors and they apply their method in the biomedical domain. These in-text citations represent the changing role of publications in a community. The authors end with an interpretatation of the proposed methods on the basis of one PubMed article from 2006.

In the paper "CitationAS: A Summary Generation Tool Based on Clustering of Retrieved Citation Content", **Jie Wang, Shutian Ma and Chengzhi Zhang** [9] focus on an automatic summary generation tool CitationAS that uses citation sentences to construct summaries. The authors build a new application which can automatically generate a summary on a given topic by optimizing the search results using clustering engine, named Carrot2, in three stages: similar cluster label merging, important sentences extraction and summary generation.

The paper "Temporal Evolution, Research Themes, and Emerging Trends in Case-Based Reasoning Literature" by **Dongxiao Gu, Bo Liu, Isabelle Bichindaritz and Changyong Liang** [4] proposes a study on the temporal evolution and emerging trends in a specific scientific field which is case-based reasoning. They analyze a dataset of 4460 papers published from 2000 to 2015. The authors study the temporal distribution of papers, the reference and journal co-citation network, and also the co-occurrence of keywords. The methodology used in this paper to provides an extensive study of a scientific field can further be applied to other fields.

The discovery of potential collaborations is in the focus of the last paper, "Mining the Potential Collaborative Relationships Based on the Author Keyword Coupling Analysis and Social Network Analysis" by **Yufang Peng,**

[1] See the proceedings of the first edition of the workshop: `http://ceur-ws.org/Vol-1384/`, [1].

[2] `https://easychair.org/cfp/CLBib2017`

Gu Dongxiao and Shi Jin [7]. Among the methods that are used are social network analysis, keyword and co-word analysis and clustering. Considering the hypothesis that authors that work on similar topics and keywords could potentially be contributors, this paper provides a method for author similarity analysis.

4 Outlook

The interest for this interdisciplinary research has been growing during the last years (see e.g. the workshops of BIRNDL - "Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries" and WoSP - "Workshop on Mining Scientific Publications") and the series of CLBib workshops up to now have shown that both fields of Natural Language Processing and Bibliometrics can benefit from addressing the problem of the full text processing of papers.

As a result of this workshop series, a new Research Topic "Mining Scientific Papers: NLP-enhanced Bibliometrics"[3] has been launched as part of the "Frontiers in Research Metrics and Analytics" journal published in Open Access. We intend to continue the effort to bring both communities together and foster the development of semantic technologies dedicated to Bibliometrics and Scientometrics.

4.0.1 Acknowledgements

Part of this research has been funded by the FEDER (Fonds européen de développement régional) and selected by the French-Swiss programme Interreg V: Webso+ project[4].

References

[1] Atanassova, I., Bertin, M., Mayr, P.: Editorial for the first workshop on mining scientific papers: Computational linguistics and bibliometrics. In: Proceedings of the First Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics co-located with 15th International Society of Scientometrics and Informetrics Conference (ISSI 2015), Istanbul, Turkey, June 29, 2015. pp. 1–4 (2015), http://ceur-ws.org/Vol-1384/editorial.pdf

[2] Bertin, M., Atanassova, I., Larivière, V., Gingras, Y.: The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology (JASIST) 67(1), 164–177 (2016), http://dx.doi.org/10.1002/asi.23367

[3] Chen, C.: Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology 57(3), 359–377 (2006), http://dx.doi.org/10.1002/asi.20317

[4] Gu, D., Liu, B., Bichindaritz, I., Liang, C.: Temporal evolution, research themes, and emerging trends in case-based reasoning literature. In: Atanassova, I., Bertin, M., Mayr, P. (eds.) 2nd Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics collocated with 16th International Conference on Scientometrics and Informetrics (ISSI 2017). CEUR-WS.org (2017)

[5] He, J., Chen, C.: Understanding the changing roles of scientific publications via citation embeddings. In: Atanassova, I., Bertin, M., Mayr, P. (eds.) 2nd Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics collocated with 16th International Conference on Scientometrics and Informetrics (ISSI 2017). CEUR-WS.org (2017)

[6] Mayr, P., Scharnhorst, A.: Combining bibliometrics and information retrieval: preface. Scientometrics 102(3), 2191–2192 (Mar 2015), https://doi.org/10.1007/s11192-015-1529-2

[7] Peng, Y., Gu, D., Jin, S.: Mining the potential collaborative relationships based on the author keyword coupling analysis and social network analysis. In: Atanassova, I., Bertin, M., Mayr, P. (eds.) 2nd Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics collocated with 16th International Conference on Scientometrics and Informetrics (ISSI 2017). CEUR-WS.org (2017)

[8] Shotton, D.: Cito, the citation typing ontology. Journal of Biomedical Semantics 1(1), S6 (Jun 2010), https://doi.org/10.1186/2041-1480-1-S1-S6

[3] https://www.frontiersin.org/research-topics/7043/mining-scientific-papers-nlp-enhanced-bibliometrics
[4] http://tesniere.univ-fcomte.fr/projet-webso/

[9] Wang, J., Ma, S., Zhang, C.: Citationas: A summary generation tool based on clustering of retrieved citation content. In: Atanassova, I., Bertin, M., Mayr, P. (eds.) 2nd Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics collocated with 16th International Conference on Scientometrics and Informetrics (ISSI 2017). CEUR-WS.org (2017)