

BIR: A time and a place to envision concepts and tools around Bibliometrics*

Marc Bertin¹[0000–0003–1803–6952] and Iana Atanassova²[0000–0003–3571–4006]

¹ Laboratoire ELICO, Université Claude Bernard Lyon 1, France
marc.bertin@univ-lyon1.fr

<http://www.elico-recherche.eu/membres/marc-bertin>

² CRIT, Université de Bourgogne Franche-Comté, France
iana.atanassova@univ-fcomte.fr

The BIR and BIRNDL workshops are a unique place where people from several different domains, Bibliometrics, Information Retrieval and more broadly Natural Language Processing can exchange ideas and work together to develop new approaches for the processing of scientific corpora.

During our PhD theses at Paris-Sorbonne university, we both worked on the linguistic analysis of corpora, in the framework of enunciative linguistics. Iana Atanassova focused on the semantic dimension of discourse in scientific corpora, and in particular for applications in Information Retrieval, and Marc Bertin focused on the discursive forms present in citation contexts. This work on citation acts, with a purely linguistic approach at the beginning, is the starting point of more than 10 years of work. The link between these two approaches and our collaboration started around the same period as the first BIR workshop in 2014. At that moment Iana worked on the development of a search engine at MyScienceWork and Marc was a post-doctoral fellow at Montreal working with Yves Gingras and Vincent Larivière [8]. The BIR and BIRNDL workshops have become regular events where we can discuss our research progress and share ideas around bibliometrics and the processing of scientific papers. Iana is proud to have been the keynote speaker of BIR 2019 [1].

Considering bibliographic references in papers, we can observe a great variety in the citation contexts but also many regularities that need to be studied and categorized in order to better understand the linguistic nature of citations and build solutions for ontology population. We have approached the study of citation contexts firstly by examining the regularities in their lexical distributions in the IMRaD structure [2] and by further characterizing the different sections of papers by Factorial Correspondance Analysis [3] and later by K-means and Hierarchical clustering methods [6]. We also studied phenomena around multiple references in the same sentence [4] and negative citations [5]. The InTeReC corpus was proposed [7] to address some of the problems in developing machine learning approaches for citation classification and we conducted a preliminary study to compare such approaches with rule-based ones [10]. Finally, we ques-

* A companion video is hosted at https://youtu.be/hLgTB0av_b0.

tioned the text spans that should be considered as citation contexts [9] in view to implementing robust machine learning classifiers. We have published this entire body of work on citation contexts in the BIR and BIRNDL workshops since 2014. Today, we are convinced that it is possible to create a model of citation contexts, and such a model is a step towards both a theory of citation and improved algorithms for recommendation systems.

The BIR community is growing and is becoming the place to be to share and discuss new and original ideas around mining scientific papers. We are very happy to be part of this family.

References

1. Atanassova, I.: Beyond metadata: the new challenges in mining scientific papers. keynote talk. In: 8th BIR 2019@ECIR. Cologne, Germany (2019)
2. Bertin, M., Atanassova, I.: A study of lexical distribution in citation contexts through the IMRaD standard. In: 1st Workshop on BIR 2014@ECIR. vol. 1143, pp. 5–12. Amsterdam, Netherlands (2014)
3. Bertin, M., Atanassova, I.: Factorial correspondence analysis applied to citation contexts. In: 2nd Workshop on BIR 2015@ECIR. vol. 1344, pp. 22–29. Vienne, Austria (2015)
4. Bertin, M., Atanassova, I.: Multiple In-text Reference Phenomenon. In: 1st Workshop on BIRNDL 2016@JCDL. vol. 19, pp. 14–22. Newark, NJ, USA (2016)
5. Bertin, M., Atanassova, I.: Weak Links and Strong Meaning: The Complex Phenomenon of Negational Citations. In: 3rd Workshop on BIR 2016@ECIR. vol. 1567, pp. 14–25. Padova, Italy (2016)
6. Bertin, M., Atanassova, I.: K-means and Hierarchical Clustering Method to Improve our Understanding of Citation Contexts. In: 2nd BIRNDL 2017@SIGIR. vol. 1888, pp. 107–112. Tokyo, Japan (2017)
7. Bertin, M., Atanassova, I.: InTeReC: In-text Reference Corpus for Applying Natural Language Processing to Bibliometrics. In: 7th BIR 2018@ECIR. Grenoble, France (2018)
8. 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* 67(1), 164–177 (2016)
9. Bertin, M., Jonin, P., Armetta, F., Atanassova, I.: Identifying the Conceptual Space of Citation Contexts using Coreferences. In: 4th BIRNDL 2019@SIGIR. vol. 2414, pp. 138–144. Paris, France (2019)
10. Perier-Camby, J., Bertin, M., Atanassova, I., Armetta, F.: A preliminary study to compare deep learning with rule-based approaches for citation classification. In: 8th BIR 2019@ECIR. vol. 2345, pp. 125–131. Cologne, Germany (2019)