Information Retrieval and Semantic Annotation of Scientific Corpora

Iana Atanassova
Centre Tesnière – CRIT, Université de Bourgogne Franche-Comté
30 rue Mégevand 25030 Besançon, France
iana.atanassova@univ-fcomte.fr

Summary

Scientific papers are highly structured texts and display specific properties related to their references but also argumentative and rhetorical structure. Natural Language Processing can be applied to efficiently explore scientific corpora and develop applications for the Semantic Web, Information Retrieval, Automatic Summarization and Bibliometrics. The organization of scientific papers typically follows a standardized pattern, the well-known IMRaD structure (Introduction, Methods, Results, and Discussion). By analysing the full text of about 80,000 papers of the PLOS corpus, we studied this structure from several different perspectives. Firstly, we performed quantitative and qualitative analyses of citations and their positions in the structure of papers. Secondly, we studied the occurrences of verbs in citation contexts and their similarities across the different sections. Finally, using sentence-based similarity metrics, we quantified the phenomenon of text re-use in abstracts with respect to the IMRaD structure. This research allowed us to establish some of the invariants of scientific papers and the results are useful for implementing novel text mining and information retrieval interfaces taking into consideration the argumentative structure of papers. More specifically, they can be considered as an important element when creating linguistic resources and rule-based methods to perform fine-grained semantic analysis of scientific papers.

Key words: Information extraction, Information retrieval, scientific papers, semantic annotation, text mining, IMRaD

Introduction

Nowadays, we can witness the emergence of a new area of study in Natural Language Processing, which is NLP-enhanced Bibliometrics, or studying the properties of scientific papers and their full text content to gain insight into the evaluation of science. At the same time, the semantic processing of scientific papers is at the heart of many other applications, such as Information Retrieval, semantic publishing, information extraction and aggregation of data from scientific corpora. This increased interest in the application of NLP methods to scientific publications is the result of several important factors:

- the ever growing availability of full text scientific corpora, as a consequence of the Open Access movement;
- the emergence of standardized formats for the representation of the full text content of scientific papers (e.g. JATS, DocBook);
- the recent developments in NLP that have resulted in a number of robust and accessible tools for versatile text processing.

In this context, many studies, workshops and evaluation sessions[1] have been initiated in the recent years, aiming to provide new methods dedicated to the processing of scientific corpora. One important point of interest is the cognitive structure of scientific production and in particular the invariants that may exist in scientific writing that can be of linguistic, discourse, structural or distributional nature.

Levels of processing

Scientific papers are subject to numerous conventions, norms and editorial requirements. They are highly structured texts and often follow a specific rhetorical structure. In experimental sciences, the IMRaD structure (Introduction, Methods, Results and Discussion) (Bertin et al., 2013) has emerged as a standardised pattern that was adopted by a majority of journals during the second half of the twentieth century.

A scientific paper can be considered as a structured document of three parts: metadata (including title, author list, publication date, keywords, abstract, etc.), full text body, and bibliography (see figure 1). Bibliometric studies traditionally focus on only the metadata and bibliography. The elements that contain information in natural language are the title, the abstract, the full text body and the bibliography. Implementing efficient information retrieval and information extraction for all these elements of scientific papers is an important step towards making scientific knowledge more accessible and helping scientists cope with the enormous amount of data produced each day. The information extraction and ontology population are intended to enrich the papers' metadata and thus define new facets for information retrieval.

Concerning the elements of the bibliography and their corresponding in-text citations, a new field of investigation has emerged that aims to characterize in-text citations according to their contexts. Many applications can be envisaged, among which the automatic summarisation (Wang and Zhang, 2017; See et al.

[1] E.g. CL-SciSumm Evaluation Task (https://www.aclweb.org/portal/content/3rd-computational-linguistics-scientific-document-summarization-shared-task), Semantic Publishing Challenge (https://2016.eswc-conferences.org/assessing-quality-scientific-output-its-ecosystem), BIRNDL (http://wing.comp.nus.edu.sg/~birndl-sigir2017/), BIR (https://www.gesis.org/en/services/events/events-archive/conferences/ecir-workshops/ecir-workshop-2017/), WOSP (https://wosp.core.ac.uk/jcdl2017/), CLBib (https://easychair.org/cfp/CLBib2017).

2017), the extraction of relations between authors and the creation of author networks, and the creation of new bibliometric indices.

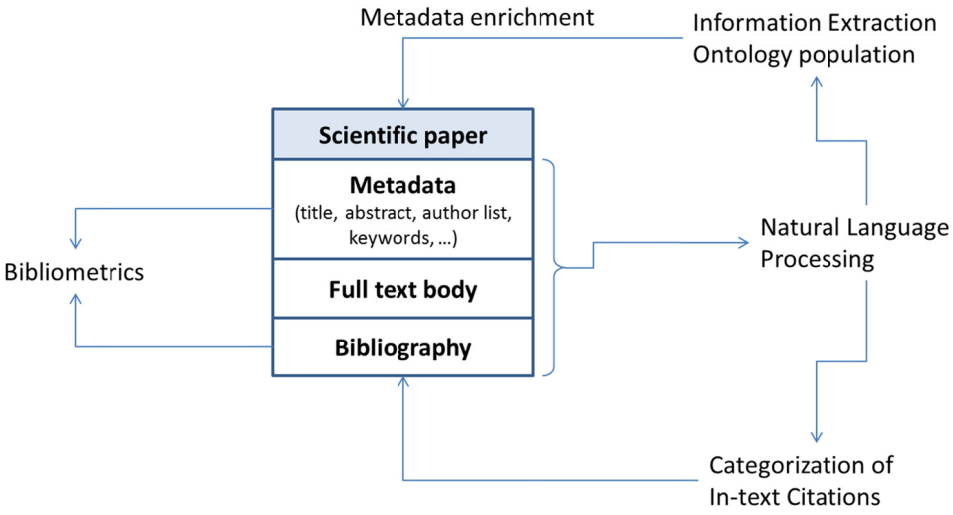

Figure 1. Elements of scientific papers and types of processing

A series of experiments on the PLOS corpus: in search of invariants

We have carried out a series of studies that show some invariants in the density of in-text citations according to their positions in the IMRaD structure (Bertin et al, 2017). The dataset consisted of around 80 000 peer reviewed papers published by PLOS[2] in Open Access. The seven PLOS journals cover different domains and are available in XML format following the JATS[3] schema. The processing stages consist of the analysis of section titles to categorize the sections, the text segmentation into sentences and the identification of in-text citations and their corresponding references. The analysis showed the existence of a strong and stable relationship between the distribution of in-text citations and the rhetorical structure of the papers, and this independently of the journals and domains (Bertin et al, 2017; 2013). This distributional phenomenon is an important factor in the study of citations and their roles in bibliometrics. Moreover, it has direct applications in information retrieval and extraction, as the density of citations in a particular sentence, paragraph or section can be related to its relevance to a user need. In fact, one possible application is the exploitation of these corpora through rich search interfaces (Bertin and Atanassova, 2012;

[2] The Public Library of Science, https://www.plos.org

[3] Journal Article Tag Suite, https://jats.nlm.nih.gov/

Presutti et al., 2014). Another related study addresses the question of the relationships that exist between abstracts, keywords and the full text body of papers (Atanassova et al., 2016). Studying the ways authors summarize their papers is useful to gain insights into which parts of the information in a paper are the most relevant and are considered as the most important by the author.

Other studies were focused on the distributional analysis of citations in IMRaD in relation to the occurrences of different verbs in the citation contexts (Bertin and Atanassova, 2014). While it is natural to find a large number of in-text citations in the introduction and literature review of an article, knowing the verbs and syntactic patterns that introduce these citations is essential to the lexical and semantic analysis of citation contexts.

The production of visualizations of scientific corpora at a large scale is important for the discovery of trends and innovations, and for landscaping a particular domain. Several visualizations of the structure of papers were produced in order to show the various roles that citations take as a function of their position in the rhetorical structure (Bertin et al., 2014). Furthermore, we have explored the visualization of spatial data extracted from corpora in the biomedical domain: we have studied the *PLOS Neglected Tropical Diseases* journal in order to produce geographical maps of the occurrences of spatial data in the corpus related to tropical diseases (Atanassova et al, 2015).

Information retrieval and semantic facets

The Semantic Web assimilates the production of scientific knowledge through ontologies dedicated to the description of scientific papers (CiTO for the characterization of citations, DoCO for the different elements of a document, and BiRO for the description of bibliographic references) (Shotton, 2010). The current publication models allow more and more applications oriented towards the exploitation of scientific corpora at a large scale and the new challenges exist around the automatic aggregation and production of semantic information related to or extracted from publications. The technologies of the Semantic Web play an important role for these tasks, especially to ensure the interoperability between the different formats and systems.

We have implemented two prototypes that use linguistic analyses combined with Semantic Web technologies: an Information retrieval system using semantic facets, and a system for the semantic processing and categorization of in-text citations (in Presutti et al., 2014), that participated in the evaluation track *Semantic Publishing Challenge* 2014 at the *European Semantic Web Conference*. The main objective is to produce data from scientific corpora in order to take into account qualitative information extraction from the papers. The annotation of the set of documents was used to produce data in the form of Linked Open Data in order to identify and characterize the cited papers, auto-citations, multiple citations of a paper, funding organizations, paragraphs and sections containing the state of the art, methods and results used in cited papers, etc. We use

knowledge-based linguistic methods for the annotation, combined with existing tools for Named Entity Recognition and POS-tagging (Chang et al., 2016; Maning, 2011). The results are presented in a semantic search engine based on SPARQL.

Facets in information retrieval are traditionally filters that are available in the user interface and that allow to refine the result list according to predefined categories present in the documents. In the classical model, these categories correspond to classes that exist in the metadata of the documents. We can go further by considering semantic facets that depend not only on metadata but also on the full text content and whose values are obtained by a linguistic analysis that must be carried out during the indexation process. Such a model gives a new possibility for the user to access the semantic content of documents: the search results can be filtered according to semantic categories and rich textual navigation can be supported using a linguistic ontology. A different way of viewing this idea is the enrichment of metadata (Bertin and Atanassova, 2012). For example, the prototype that we have implemented gives the possibility for search and textual navigation in a scientific corpus at the level of the sentence by the choice of categories present in the papers, such as *method, hypothesis, result, conclusion, opinion,* etc. These semantic categories are identified during the indexing by a knowledge-based approach. The interface allows a semantic search, where the relevance of a sentence is a function of both keywords and semantic relations expressed in the sentence.

Conclusion

We have presented an overview of several applications around the exploitation of scientific corpora: a distributional study and the characterizing of in-text citations, a semantic search engine and spatial data visualization. With the growing number of papers published daily scientists need new tool for more efficient access to the information and to be able to rapidly grasp the main ideas of papers. The development of such tools is a recent area of research and has been greatly favored by the Open Access movement. One major challenge before Natural Language Processing is modelling and automating the extraction of argumentative structures for the construction of new textual navigation interfaces for scientific texts.

References

Atanassova, Iana; Bertin, Marc; and Larivière, Vincent. On the Composition of Scientific Abstracts. Journal of Documentation. 72 (2016), 4; pp.636 – 647

Atanassova, Iana; Bertin, Marc and Kauppinen, Tomi. Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales, Gestion et Analyse des données Spatiales et Temporelles (GAST'2015), 15ème conférence internationale sur l'extraction et la gestion des connaissances (EGC-2015), Luxembourg.

Bertin, Marc ; Atanassova, Iana; Larivière, Vincent and Gingras, Yves. The Invariant Distribution of References in Scientific Papers. Journal of the Association for Information Science and Technology (JASIST). 17 (2017), 1, pp. 164 – 177

Bertin, Marc and Atanassova, Iana. A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. Bibliometric-enhanced Information Retrieval Workshop at the 36th European Conference on Information Retrieval (ECIR-2014), Amsterdam, Netherlands.

Bertin, Marc ; Atanassova, Iana; Larivière, Vincent and Gingras, Yves. The Linguistic Context of Citations. International exposition: 10th Iteration of the Places & Spaces: Mapping Science Exhibit on "The Future of Science Mapping", 2014

Bertin, Marc ; Atanassova, Iana; Larivière, Vincent and Gingras, Yves. The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. 14th International Society of Scientometrics and Informetrics Conference (ISSI-2013), Vienna, Austria.

Bertin, Marc and Atanassova, Iana. Semantic Enrichment of Scientific Publications and Metadata: Citation Analysis Through Contextual and Cognitive Analysis. 18, 7-8. *D-lib Magazine*. 2012

Chang, Angel; Spitkovsky, Valentin I.; Manning, Christopher D. and Agirre, Eneko. A comparison of Named-Entity Disambiguation and Word Sense Disambiguation. International Conference on Language Resources and Evaluation (LREC 2016).

Manning, Christopher D. Part-of-Speech Tagging from 97\ to 100\: Is It Time for Some Linguistics? Conference on Intelligent Text Processing and Computational Linguistics (CICLing). 2011

Presutti, V., Stankovic, M., Cambria, E., Cantador, I., Di Iorio, A., Di Noia, T., Lange, C., Reforgiato Recupero, D., Tordai, A. (Eds.), SemWebEval 2014 at ESWC 2014, *Semantic Web Evaluation Challenge*, Communications in Computer and Information Science (Book 475), Springer, Anissaras, 2014

See, Abigail; Liu, Peter J and Manning, Christopher D. Get To The Point: Summarization with Pointer-Generator Networks. Association of Computational Linguistics (ACL), 2017

Shotton, D. (2010). CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics, 1 (Suppl 1), S6. doi:10.1186/2041-1480-1-S1-S6

Wang, Jie and Zhang, Chengzhi. CitationAS: A Summary Generation Tool Based on Clustering of Retrieved Citation Content. 2nd Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics at ISSI 2017, Wuhan, China.