Beyond Metadata: the New Challenges in Mining Scientific Papers

Iana Atanassova[0000−0003−3571−4006]

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

Abstract. Scientific articles make use of complex argumentative structures whose exploitation from a computational point of view is an important challenge. The exploration of scientific corpora involves methods and techniques from Natural Language Processing in order to develop applications in the field of Information Retrieval, Automatic Synthesis, citation analyses or ontological population. Among the problems that remain to be addressed in this domain is the developing fine-grained analyses of the text content of articles to identify specific semantic categories such as the expression of uncertainty and controversy that are an integral part of the scientific process. The well-known IMRaD structure (Introduction, Methods, Results, and Discussion) is often used as a standard template that governs the structure of articles in experimental sciences and provides clearly identifiable text units. We study the internal structure of articles from several different perspectives and report on the processing of a large sample extracted from the PLOS corpus. On the one hand, we analyse citation contexts with respect to their positions, verbs used and similarities across the different sections, and on the other hand, we study other phenomena such as the expression of uncertainty. The production of standard datasets dedicated to such tasks is now necessary and would provide favourable environment for the development of new approaches, e.g. using neural networks, that require large amounts of labelled data.

CEUR-WS.org/Vol-2345/paper1.pdf

1 Introduction

Scientific papers have evolved as the major communication vector to diffuse state-of-the-art knowledge and new findings among the scientific community. The development of efficient data mining tools dedicated to the information retrieval and information extraction of papers is related to our capacity to analyse the text content of papers through Natural Language Processing (NLP). Considering the structural elements of a paper, we can distinguish at least three different types of data:

– its metadata, i.e. title, author list, journal or conference information, abstract and keywords;
– the body of the paper, composed of structured text and other non-textual elements (figures, tables, etc.);

- the bibliography, which, together with the in-text citations, relates the paper to other existing research.

To convey new knowledge, scientific articles make use of complex argumentative structures whose exploitation from a computational point of view is a challenging task. The practices over time have established various frameworks used by researchers to structure their discourse. One such framework is the IMRaD structure (Introduction, Methods, Results and Discussion) which has become predominant in experimental sciences where it is considered as the outcome of the evolution of scientific publishing [21, 17].

2 Studying the Internal Structure of Papers

One of the major advantages of this IMRaD structure is the fact that it provides clearly identifiable text units, making it possible to read only parts of an article and access rapidly sections that correspond to a specific information need. Indeed, the organization of papers in the IMRaD format often result from editorial requirements and rules that are followed by the authors when producing a paper [20, 15]. The objective is to provide papers with uniform structures, thus facilitating access to information.

In recent years, several studies on the internal structure of articles, and more specifically the IMRaD structure, from the point of view of bibliometrics have been carried out. In 2013 we proposed a first study of the distribution of in-text references in the different sections of IMRaD of about 80,000 papers published in PLOS [5, 6]. This first experiment showed that the density of in-text citations per sentence is strongly dependent on the position in the section and the section type, and the distribution follows a specific pattern, the Introduction and Discussion sections containing the largest number of references.

A different perspective was taken in the study of verb occurrences in citation contexts with relation to their positions in IMRaD [4]. The verbs that appear in citances, i.e. sentences containing references, are likely to indicate the relationship that exists between the citing paper and the cited work. More generally, the study of the syntactic patterns that introduce the in-text references is important for the classification of citation contexts. The results of this experiment indicate that the ranked lists of verbs for the fours different sections differ considerably, and the Methods section makes use of verbs which are, for the large part, rare in the other sections. We further studied the linguistic patterns found in citation contexts based on the frequency of n-gram co-occurrences to identify the function of citations [7].

In the past two years, several other studies have shown interest in the IMRaD structure of papers and its relation of in-text references from various perspectives and using different datasets.

Analysing the databases of PubMed Central Open Access subset and Elsevier journals, [8] produce the distributions of in-text references depending on textual progression and scientific fields. They report on the fact that reference counts depend strongly on the scientific field. Also considering PubMed Central Open Access, [18] examines the possibility to differentiate between articles based on the citation counts that originate from different sections. The study shows that the sections of IMRaD cannot be considered as reliable indicators of citation context by themselves.

Using the IMRaD structure to refine the traditional bibliographic coupling, [10] shows that distinguishing between the citations in the different sections helps to improve paper recommendation.

In a case study of the citations of the retracted papers J. H. Schön [14], a well-known example of scientific misconduct, the authors examine the sections in which these references appear and conclude that the retracted papers are cited for the largest part in the Introduction section, while half of the citations of the non-retracted papers are in non-Introduction sections. This result indicates a there exists some measurable relationship between the position of an in-text reference and the quality of the cited research.

When studying the internal structure of papers, one important point is the structure of the abstract and its relationship to parts of the body of the paper. In fact, analysing the positions of the text segments that authors naturally use to produce a summary of their paper is useful to better understand the IMRaD structure and the positions in the textual progression where the most important information is found. We studied the distributions of sentences in the body of the papers that are reused in abstracts with respect to the IMRaD structure [2] and found out that the beginning of the Introduction and the second part of the Discussion section contain the sentences that produce the largest part of the abstract, while the Methods and Results sections contribute less. Similar results were reported by another study[19], which proposes to measure the similarity among words and paragraphs by using network architecture for word and paragraph embeddings (Doc2Vec). This study shows that paragraphs in the Introduction and the Discussion sections are more similar to the abstract than the rest of a paper and the Methods section is least similar to the other sections.

3 Uncertainty Mining

By definition, papers that report on novel research contribute to the current state of the art by adding new ideas and findings. In this respect, on tasks that seem important is the possibility to landscape the hypotheses and research topics and remain yet to be investigated in a given field. The existence of such research topics that have already been identified by researchers can be signalled in papers by various expressions of modality and uncertainty, e.g. *"these findings provide evidence that..."*, *"these results could be further used for ..."*, *"more experiments are needed to ..."*, etc.

While uncertainty mining in general has been the subject of many studies, several experiments were conducted on scientific corpora, and in particular with the aim to study the structure of papers.

We examined [3] two datasets of papers in the fields of Biomedicine and Physics from PubMed Central Open Access, and analysed the positions in the IMRaD structure where uncertainty is likely to be expressed. We compare the distributions of simple cue words and linguistic expressions that we call strong indicators of uncertainty. The results show that authors express uncertainty mostly in the Discussion section and in general towards the end of the textual progression. We observe significant differences between the two fields.

In a different approach, [16] classify a set of biomedical papers by method, type of method and non-methods by examination of citation contexts and using supervised machine learning. This study showed that hedging words play an important role for non-methods, and further that hedging is inversely related to citation frequency.

4 New Challenges and Perspectives

The research around mining scientific papers has developed rapidly in recent years thanks to the Open Access movement, which has provided large corpora of papers, and the recent advances in Natural Language Processing that have made possible scaling up the processing of texts, especially for the tasks related to the categorization of text segments.

When defining the future research directions in this field, we should ask one major question: when dealing with papers what are the tasks that we should try to automate and to what extent? In other words, what are the automatic processing tools that could facilitate access to scientific knowledge and help structure the available information without limiting the creativity and the capacity of knowledge discovery for researchers (see e.g. [13])?

Among the problems that remain to be addressed in this domain is the development of fine-grained analyses of the text content of articles to identify specific semantic categories such as the expression of uncertainty and controversy. In fact, these two categories are an integral part of the scientific process [11, 12]. The current methods developed in these fields provide only partial answers, e.g. by extracting hypotheses or speculations (see e.g. [9]).

Recently, the development of deep neural networks in the field of NLP has given good results for a variety of tasks around text mining and classification [1]. The quality of the trained models is strongly dependent on the size and quality of the datasets. Obtaining annotated datasets that require manual annotation and/or checking is costly. For this reason, one important factor that could foster the development of new methods in mining scientific papers is the availability of large annotated corpora that could be used as shared gold standards by the community. The definition of the annotation categories of such corpora is of utmost importance, as their granularity and well-foundedness define the possible applications the usefulness of such corpora. In this respect, the annotation schemas need to be designed in a way to enable the reproducibility of the categorization for other datasets.

References

1. Altınel, B., Ganiz, M.C.: Semantic text classification: A survey of past and recent advances. Information Processing & Management **54**(6), 1129 – 1153 (2018). https://doi.org/https://doi.org/10.1016/j.ipm.2018.08.001
2. Atanassova, I., Bertin, M., Larivière, V.: On the composition of scientific abstracts. Journal of Documentation **72**(4), 636 – 647 (Jul 2016). https://doi.org/10.1108/jdoc-09-2015-0111
3. Atanassova, I., Rey, F.C., Bertin, M.: Studying Uncertainty in Science: a distributional analysis through the IMRaD structure. In: WOSP - 7th International Workshop on Mining Scientific Publications at 11th edition of the Language Resources and Evaluation Conference. Miyazaki, Japan (5 2018)
4. Bertin, M., Atanassova, I.: A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. In: Bibliometric-enhanced Information Retrieval Workshop at the 36th European Conference on Information Retrieval (ECIR). Amsterdam, Netherlands (4 2014)
5. Bertin, M., Atanassova, I., Larivière, V., Gingras, Y.: The distribution of references in scientific papers: An analysis of the IMRaD structure. In: 14th International Conference of the International Society for Scientometrics and Informetrics. Proceedings of ISSI 2013, vol. 1, pp. 591–603. Vienne, Austria (Jul 2013)
6. Bertin, M., Atanassova, I., Larivière, V., Gingras, Y.: The invariant distribution of references in scientific papers. Journal of the Association for Information Science and Technology (JASIST) (May 2015). https://doi.org/10.1002/asi.23367
7. Bertin, M., Atanassova, I., Sugimoto, C., Larivière, V.: The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. Scientometrics **109**(3), 1417 – 1434 (Dec 2016). https://doi.org/10.1007/s11192-016-2134-8
8. Boyack, K.W., van Eck, N.J., Colavizza, G., Waltman, L.: Characterizing in-text citations in scientific articles: A large-scale analysis. Journal of Informetrics **12**(1), 59–73 (2018)
9. Díaz, N.P.C., López, M.J.M.: Negation and Speculation Detection, vol. 13. John Benjamins Publishing Company (2019)
10. Habib, R., Afzal, M.T.: Sections-based bibliographic coupling for research paper recommendation. Scientometrics (Mar 2019). https://doi.org/10.1007/s11192-019-03053-8
11. Jamieson, D.: Scientific uncertainty and the political process. The ANNALS of the American Academy of Political and Social Science **545**(1), 35–43 (1996). https://doi.org/10.1177/0002716296545001004
12. Kamhi, A.G.: Balancing certainty and uncertainty in clinical practice. Language, Speech, and Hearing Services in Schools **42**(1), 59–64 (2011). https://doi.org/10.1044/0161-1461(2009/09-0034)
13. Kotkov, D., Veijalainen, J., Wang, S.: Challenges of serendipity in recommender systems. In: WEBIST 2016: Proceedings of the 12th International conference on web information systems and technologies. vol. 2. SCITEPRESS (2016)
14. Luwel, M., van Eck, N.J., et al.: The Schön case: Analyzing in-text citations to papers before and after retraction. In: 23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands. Centre for Science and Technology Studies (CWTS) (2018)
15. Nair, P.K.R., Nair, V.D.: Organization of a Research Paper: The IMRAD Format, pp. 13–25. Springer International Publishing, Cham (2014)

16. Small, H.: Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. Journal of Informetrics **12**(2), 461–480 (2018)
17. Sollaci, L.B., Pereira, M.G.: The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. Journal of the medical library association **92**(3), 364 (2004)
18. Thelwall, M.: Should citations be counted separately from each originating section. arXiv preprint arXiv:1903.07547 (2019)
19. Thijs, B.: Paragraph-based intra-and inter-document similarity using neural vector paragraph embeddings. FEB Research Report MSI_1901 (2019)
20. Todorović, L.: Original (scientific) paper: The imrad layout. Archive of Oncology **11**(3), 203–205 (2003)
21. Wu, J.: Improving the writing of research papers: Imrad and beyond. Landscape Ecology **26**(10), 1345–1349 (Dec 2011). https://doi.org/10.1007/s10980-011-9674-3