Characterizing In-text Citations using N-gram Distributions

Marc Bertin[1] and Iana Atanassova[2]

[1] bertin.marc@gmail.com
Centre Interuniversitaire de Rercherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal (UQAM), Canada

[2] iana.atanassova@univ-fcomte.fr
Centre Tesniere, University of Franche-Comte, France

Introduction

This article focuses on a Natural Language Processing (NLP) approach for the analysis of citation functions in scientific papers. Bibliometric studies traditionally rely on citation metadata and count the number of times a publication has been cited. However, some recent studies rely also on full text processing on papers, e.g. (Boyack et al., 2013), (Bertin et al., 2013, 2014). The full text content of papers and more specifically the sentences containing citations provide valuable information on the functions of citations that can be exploited through NLP. To study citation acts, we need to consider full text papers and their rhetorical structure.

The main question that we want to answer here is whether the most frequent citation patterns are correlated to the rhetorical structure of scientific papers. We investigate the properties of the linguistic patterns that appear in citation contexts. For this, we study the distribution of n-gram classes containing verb forms, and we show the existence of three different types of distributions according to the rhetorical structure.

Method

By analyzing a large corpus of articles, we propose a quantitative study of the linguistic patterns around in-text citations. Some words or sets of words in n-grams are more frequent than others (Cavnar & Trenkle, 1994), and this idea is consistent with Zipf's Law (Zipf, 1949). The difficulty is that the calculation of n-grams in contexts results in a combinatorial explosion. We propose several filters to reduce the number of patterns.

The rhetorical structure of scientific papers is typically organized around a standardized pattern, known as the IMRaD structure (Introduction, Methods, Results and Discussion). We identify the four main section types of this structure by analysing section titles. Then, we consider the set of sentences containing citations and belonging to each section type.

We represent citation contexts by using sequences of words of length n called n-grams where $2<n<=5$. In our approach we consider only n-grams within sentence boundaries because sentences are natural building blocks of the text. For each n-gram we observe its frequencies in the four section types of the IMRaD structure. For our study, we select only the n-grams that contain at least one verb form. In this way, the number of n-grams to process is much smaller and we eliminate word patterns containing only nominal groups like: “In this paper”, “the present article”, “the result of” etc. for 3-grams.

Dataset

We performed an automatic analysis of the seven peer-reviewed academic journals published in Open Access by the Public Library of Science (PLOS). The corpus contains about 85,660 research articles. Most of the articles are in the biomedical domain, but the corpus covers all fields of Human and Natural Sciences, as the publisher’s main journal, PLOS ONE, is multidisciplinary. Around 98% of the articles in the corpus follow the IMRaD structure, which is imposed by editorial requirements.

Results

We select the most frequent verb forms in order to construct n-gram classes from in-text citation contexts. This data will be used to obtain a first typology of the distribution of n-grams depending on the rhetorical structure of articles.

The following figures present distributions of n-grams classes for the IMRaD sections. We can distinguish between three different type of classes, and we give one example of each. The horizontal axis presents the text progression of the section from 0% to 100%. The vertical axis gives the percentage of occurrences of each class relative to its occurrences in citation contexts in the entire article.

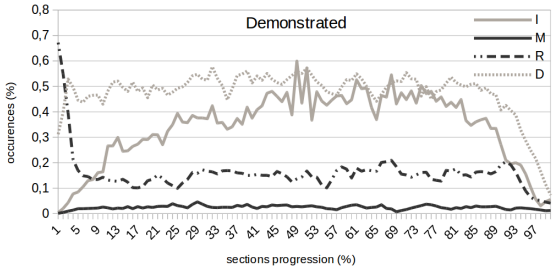

Figure 1. Demonstrated.

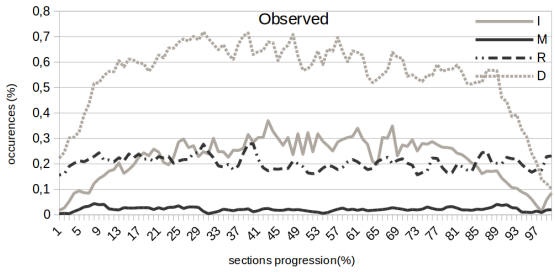

Figure 2. Observed.

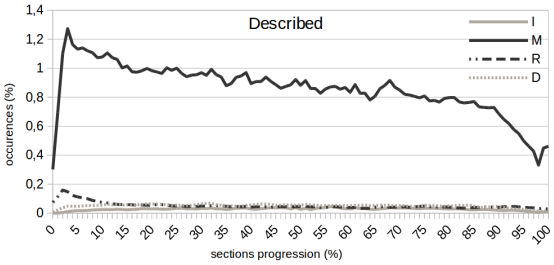

Figure 3. Described.

Discussion

Figure 1 shows the first class, which includes n-grams containing the verb *Demonstrated***.** These n-grams appear with roughly equivalent frequencies in the sections Results and Discussion, but, at the same time the Methods section contains much lower frequencies of these patterns.

Figure 2 shows the second class type, which includes n-grams with the verb *Observed***.** We can observe another type of distribution, with relatively very high frequencies in the Discussion section.

Figure 3 shows the distribution of n-grams with the verb *Described*. We can observe that the structure of the Methods section is unique, as the class *Described* is present with a very high frequency in this section and especially at the beginning of the section. Moreover, Figures 1 and 2 show that on the distributions for the other classes, the Methods section contains relatively few occurrences. In other words, the class *Described* is characteristic of the Methods section, where it appears with very high frequency, and it is very rare in all the other sections. The Methods section displays very low frequencies for all classes except *Described*.

These results imply that each section, depending on its nature, authorizes more or less easily the usage of specific patterns containing verbs. The Methods section is rather closed in nature, where we find a very small number of high frequency verbs. At the same time, the Discussion section is open to different forms and allows a larger number of variations in terms of the linguistic means that authors use in citation contexts.

Conclusion

The purpose of this study is to demonstrate the existence of frequent n-gram patterns in citation contexts and their strong relation with the rhetorical structure of scientific articles. Studying the n-gram classes containing verb forms, we show the existence of three different types of distributions according to the rhetorical structure. From our point of view, the problem of the automatic annotation of citation contexts is strongly related to identifying significant surface patterns for the annotation process.

Acknowledgments

We thank Benoit Macaluso of the Observatoire des Sciences et des Technologies (OST), Montreal, Canada, for harvesting and providing the PLOS data set.

References

Bertin, M., & Atanassova, I. (2014). A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval* (ECIR 2014). Amsterdam, The Netherlands.

Bertin, M., Atanassova, I., Larivière, V., & Gingras, Y. (2013). The Distribution of References in Scientific Papers: an Analysis the IMRaD Structure. *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics* (pp. 591–603). Vienna.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology, 64*(9),1759–1767.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. Ann Arbor MI, 48113(2), 161–175.

Small, H. (1982). Citation context analysis. *Progress in Communication Sciences, 3*, 287–310.

Zipf, G. K. (1949). Human behavior and the principle of least effort.