Identifying the conceptual space of citation contexts using coreferences

Marc Bertin[1][0000−0003−1803−6952], Pierre Jonin[2], Frédéric Armetta[2], and Iana Atanassova[3][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] Laboratoire LIRIS, Université Claude Bernard Lyon 1, France
pierre.jonin@etu.univ-lyon.fr - frederic.armetta@univ-lyon1.fr
[3] CRIT, Université de Bourgogne Franche-Comté, France
iana.atanassova@univ-fcomte.fr

Abstract. The study of citation contexts is an important element in understanding the function of citations and categorizing the relationships between works. One of the problems in this field is defining the size of citation contexts. In this paper we propose the definition of citation blocks (CB) that are citation contexts composed of one or more sentences that are linked by coreference clusters. We describe the methodology for the automatic processing and determining the boundaries of CB and observe the different sizes of CB in the different sections of the IMRaD structure of articles. The results are obtained from a sample of 70,000 citation contexts extracted from the PLOS dataset.

Keywords: Citation contexts · Coreference Resolution · Deep learning.

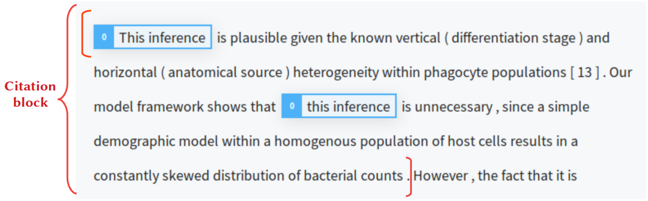
1 Introduction

In the bibliometric field, work on the study of the full text of articles offers both fascinating perspectives and technological and conceptual limitations. While bibliographic metadata are well structured, dealing with text implies more complexity in the processing of unstructured data. In many cases it is necessary to delimit certain areas in the text in order to process the information contained in these areas. A lot of research is based on the identification of textual spaces and argumentative zones e.g. (see Teufel [18] and [17]). From the point of view of the rhetorical structure of articles, the IMRaD structure has already been studied by [1]. Understanding citation acts is a necessary step in the categorization of semantic relationships between works. The access to the full text of articles is an essential step in this process and the Open Access and Open Science movements play are favourable for the development of such approaches. Indeed, if Peroni & Shotton (see [13]) have already proposed an ontological modelling of references, bibliometricians are still looking for a theory of citation (see Cronin [3]). The study of the full text of articles can provide relevant empirical results and contribute to the ontological population of these semantic web models. Using citation contexts is one of the tasks that is important to understand and predict the behaviour of citation acts. For exemple, [16] analyze a dataset of 1.5 million computer science articles and more than 26 million citation contexts to extract some feature to predict long-term citation behavior. Other research (see [7]) shows that authors are sensitive to discourse structure. They analyze how scientific works frame their contributions through different types of citations by introducing a new dataset of nearly 2,000 citations annotated for their function. Kaplan [8] has built a corpus of 38 articles from Computational Linguistics: Special Issue on the Web as Corpus. The purpose of their work is to underline the importance of using citation contexts for the synthesis of research documents.

To this end, we will propose a method for determining textual spaces that facilitate the study of citation contexts in order to produce both qualitative and quantitative analyses. To do this, we must focus on the concept of anaphora, cataphora and deixis, as well as co-references in order to propose a methodology to define meaningful textual spaces for the analysis of citation contexts. In recent years, two approaches have coexisted in the treatment of citation contexts. The first one is based on the segmentation of the text into sentences which, from a linguistic point of view, represent a unit of meaning [1]; the second approach is based on the choice of the size of a window, variable or not, which will determine the context around an in-text reference [2, 15, 17, 14, 4]. It is true that from a quantitative point of view, choosing a window size involves a risk of overlap. However, neither approach is satisfactory. The first generates a form of indeterminacy, the second induces an intrinsic noise. The choice of one or the other approach is generally motivated by the nature of the tools and methods use, as well as the desired result. An in-depth study of the context of citations shows that the argumentation developed in the elements of the discourse by an author, according to the rhetorical structure of an article, can introduce variations in the semantic value of the relationship. We propose to address this issue through the study of anaphoric relationships and co-references.

2 Research Problem

The treatment of semantic-pragmatic phenomena such as anaphora, cataphora and deixis is of central importance to us in the analysis and categorization of citation acts. We will therefore define these notions in order to show their importance, and also, by the nature of these relationships, the existence of a space where these relationships can be expressed. We face a problem of co-references and anaphorical relations. Halliday and Hasan ([6]) use the notion of cohesion to define the nature of the anaphoric relationship. A referential object is called an anaphora when it refers to its antecedent. It may be a previously introduced expression but does not necessarily designate the same entity as that expression. The anaphora may be grammatical, lexical, nominal or pronominal in nature, but also adverbial, verbal, summarizing, associative, etc., underlining the complexity of this phenomenon. The process of searching for this precedent is generally referred to as 'anaphora resolution'. A co-reference can be defined as a reference to the same entity whose context alone can establish the link between the two expressions. This can lead to the successive identification of corefential chains (see Mitkov 1989). We also face a problem of contextual and co-referential space. From a linguistic point of view, Kleiber [9] refers to "the immediate environment" as the "linguistic context" for anaphors and "the immediate denunciation situation" for deictics (see [12]). The consideration of a space based on the work around the anaphors makes sense and can provide a solution to the limitations presented by [14, 15].



0 This inference is plausible given the known vertical (differentiation stage) and horizontal (anatomical source) heterogeneity within phagocyte populations [13] . Our model framework shows that 0 this inference is unnecessary , since a simple demographic model within a homogenous population of host cells results in a constantly skewed distribution of bacterial counts . However , the fact that it is

Fig. 1. Coreference clusters and citation blocks

This problem leads us to hypothesize that the space of citation contexts must be extended beyond the sentence and within a space delimited by criteria of a semantic/linguistic nature and not quantitative, i.e. according to a window delimited by numerical values. To do this, we propose to study for this paper co-referential relationships in order to determine their presence or not, and if so, what is the size of this co-referential space. As an example, figure 1 shows the textual space around an in-text reference. The expressions "this inference" belong to a coreference cluster and are in the first and second sentence. The citation block (CB) is thus delimited by these two sentences.

3 Method

For the following experiment, we have taken as example a corpus of 70,000 in-text references extracted from articles published by PLOS. Our dataset contains in-text references and their contexts chosen randomly from each of the 7 PLOS journals (10,000 citation contexts from each journal). The articles were downloaded in XML format and for each in-text reference we have extracted the metadata related to the article, the metadata related to the cited work, the title of the section, as well as the paragraph containing the in-text reference that will be further processed for the identification of the citation block. The section titles follow, for the large majority of the articles, the IMRaD structure (Introduction, Methods, Results and Discussion). This structure for the PLOS corpus has already been extensively studied in relation with in-text references (see [1]).

We have processed the section titles to identify the section type for each in-text reference and in the rest of the article the sections are coded with the letters I, M, R and D. A small portion of the section titles did not follow this pattern and could not be classified as I, M, R and D. They were excluded from the sample. From an implementation point of view, we have used the latest advances in the field of co-referencing. We annotated the corpus from the allenNLP libraries [5] dealing with co-references and based on the end-to-end coreference resolution model [10]. This model surpasses all previous work with two specific points: they do not use a syntax analyzer and they do not use a manual mention detector. For each in-text reference, we first identify the textual space (TS) that can possibly be related to the reference through the use of coreference and anaphoric expressions in the following way: TS is composed of the sentence containing the in-text reference and all the following sentences until a new in-text reference is encountered within the same paragraph. In fact, we consider two types of boundaries that delimit the TS related to in-text references: paragraph breaks and the presence of other references. In fact, when a new in-text reference is encountered in a paragraph, we suppose that the sentences immediately following this reference could be related to it, provided that they do not contain other in-text references. To identify all textual elements that belong to coreference clusters in the sentences we have used the python library AllenNLP , which implements coreference resolution using the method described in [10]. Coreference clusters are sets of text elements, that can be words or sequences of words. The elements of a coreference cluster can belong to the same sentence or to different sentences. In the later case, the coreference cluster establishes a link between these different sentences. We consider that sentences that contain elements of the same coreference cluster should belong to the same citation block (CB). Given an in-text reference and its TS, we consider that the beginning of the CB is the sentence containing the in-text reference and the end of the CB is the last sentence in TS that is linked to this first sentence by the coreference clusters. In the case when the TS is composed of only one sentence, there is no need to identify the coreference clusters as the citation block is also composed of one sentence.

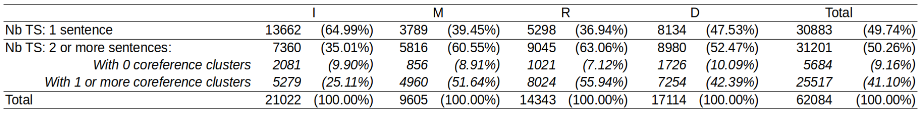
4 Results

Our aim is to observe the trends in the different journals and section types of the IMRaD structure. Table 1 presents the average sizes in sentences of TS, as well as the percentage of TS having only one sentence. We observe that the sizes of TS are relatively small for all introduction sections (between 1.43 and 1.97 sentences) and much larger for the results and methods sections (between 2.52 and 3.07). This result is consistent with the fact that in-text references are less frequent in the results and methods sections, as observed by [1]. The last column of this table gives the percentage of TS that contain only one sentence. These TS are for the most part in the introduction section and account for about half of all TS. In our approach, when the TS has only one sentence, the citation block has the same size and no analysis of the coreference clusters in necessary.

Table 1. Sizes of TS in the different journals and section types

Journal	Introduction		Method		Result		Discussion	
	⋆	†	⋆	†	⋆	†	⋆	†
pbio	1.59	61.53%	2.53	42.52%	2.76	42.52%	2.15	47.91%
pcbi	1.97	54.21%	2.52	41.76%	2.61	39.10%	2.42	37.46%
pgen	1.62	64.27%	2.80	36.70%	2.71	37.54%	2.17	47.46%
pmed	1.57	68.08%	2.78	35.46%	2.68	39.55%	1.93	52.80%
pntd	1.43	73.43%	2.67	37.85%	2.67	38.73%	2.10	45.68%
pone	1.60	66.92%	2.49	42.04%	2.84	38.88%	2.05	50.61%
ppat	1.52	68.33%	2.94	37.81%	3.07	31.95%	2.05	47.00%
⋆: Sentences in TS (average)								
†: Percentage of TS with 1 sentence								

Table 2 presents the numbers and percentages of TS with 1 sentence (49.74%) and with two or more sentences. The latter are divided in two groups: TS without coreference clusters (9.16%) and TS with 1 or more coreference clusters (41.10%). The further analysis will be done on the TS with 1 or more coreference clusters in order to delimit the citation blocks (CB) and evaluate the difference in the size of TS and CB that we obtain.



	I		M		R		D		Total	
Nb TS: 1 sentence	13662	(64.99%)	3789	(39.45%)	5298	(36.94%)	8134	(47.53%)	30883	(49.74%)
Nb TS: 2 or more sentences:	7360	(35.01%)	5816	(60.55%)	9045	(63.06%)	8980	(52.47%)	31201	(50.26%)
With 0 coreference clusters	*2081*	*(9.90%)*	*856*	*(8.91%)*	*1021*	*(7.12%)*	*1726*	*(10.09%)*	*5684*	*(9.16%)*
With 1 or more coreference clusters	*5279*	*(25.11%)*	*4960*	*(51.64%)*	*8024*	*(55.94%)*	*7254*	*(42.39%)*	*25517*	*(41.10%)*
Total	21022	(100.00%)	9605	(100.00%)	14343	(100.00%)	17114	(100.00%)	62084	(100.00%)

Table 2. Numbers of TS with 1 sentence and 2 or more sentences in IMRaD

Figure 2 presents the final result which is the average sizes of TS and CB for the different section types of the IMRaD structure. We observe that on average CB tend to be smaller than TS by about 1 sentence, and this for all sections. In fact, the sizes of CB vary between 1.79 and 2.88 sentences, the largest size being in the results section. Table 2 and Figure 2 are there to show the importance of the phenomenon of co-references in the study of citation contexts. This study could have been a negative study if the anaphoric relationships would have been weak, which is not the case and the phenomenon cannot be ignored.

5 Discussion and Conclusion

The methodological and conceptual limitations of this study are the nature of the co-reference resolution tools, which must be finer and offer more detailed analyses. Indeed, we will not discuss here the case of deictics, which is nevertheless essential. [12]) From a linguistic point of view, it would be necessary, on the one hand, to extend the discussion around this "place of existence" of the

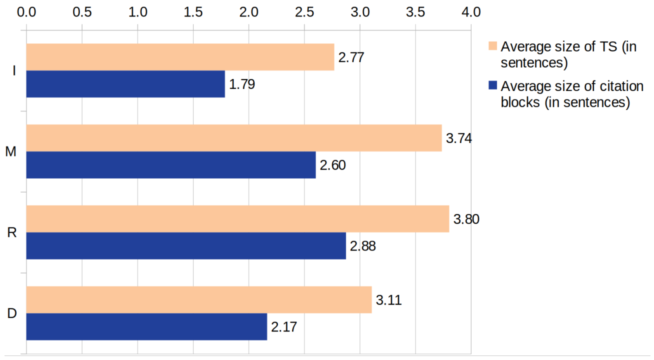


Fig. 2. Average sizes of TS and CB in sentences with respect to IMRaD

reference, to the study of endophores, namely anaphora and cataphora, from the point of view of the reference when it refers to a statement that may be a segment (segmental) or a longer statement (resomptive) [11]. This citation block model should eventually make it possible to better understand the nature of citation acts, to have a consensus on the spaces that carry information for the semantic categorization of citation contexts and to propose finer corpora dedicated to this task. It will be necessary to differentiate between anaphoric relationships and co-references. The processing operations will not be the same from the linguistic point of view and the implementation of automatic text processing tools, in order to achieve a better understanding of the mechanisms of citation acts. This paper presents an original approach in the sense that it invites us to take this linguistic phenomenon into account as a basis for the study of citation contexts. This study must be considered as a necessary and not sufficient solution. For this study to be complete, it is still necessary to determine the accuracy perimeter of End-to-end Neural Coreference Resolution Model approaches. Today, this is a promising approach that removes a technological lock around the identification of co-references. Without a mature solution around the identification of co-references and especially without evaluation, this type of model will remain conceptual. However, it seems that this type of modeling is promising in the sense that the information characterizing the citation contexts is finally defined in a space now identified. The perspectives of this work around this work focus on the problems of identifying co-references and anaphoric relationships with shallow neural networks. It is also necessary to evaluate and improve this identification by proposing learning dataset to dedicate this task to the specific processing of scientific articles.

6 Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Bertin, M., Atanassova, I., Gingras, Y., Larivière, V.: The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology **67**(1), 164–177. https://doi.org/10.1002/asi.23367
2. Bradshaw, S.: Reference directed indexing: Redeeming relevance for subject search in citation indexes. In: International Conference on Theory and Practice of Digital Libraries. pp. 499–510. Springer (2003)
3. Cronin, B.: The need for a theory of citing. Journal of documentation **37**(1), 16–24 (1981)
4. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., Zhai, C.: Content-based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology **65**(9), 1820–1833 (2014)
5. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., Zettlemoyer, L.: Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640 (2018)
6. Halliday, M.: Hasan. r.(1976). cohesion in english. L ondon: L ongman (1980)
7. Jurgens, D., Kumar, S., Hoover, R., McFarland, D., Jurafsky, D.: Measuring the evolution of a scientific field through citation frames. Transactions of the Association for Computational Linguistics **6**, 391–406 (2018)
8. Kaplan, D., Iida, R., Tokunaga, T.: Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL). pp. 88–95 (2009)
9. Kleiber, G.: Anaphore-deixis: où en sommes-nous? L'information grammaticale **51**(1), 3–18 (1991)
10. Lee, K., He, L., Lewis, M., Zettlemoyer, L.: End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045 (2017)
11. Maillard, M.: Essai de typologie des substituts diaphoriques (supports d'une anaphore et/ou d'une cataphore). Langue française (21), 55–71 (1974)
12. Perdicoyanni-Paléologou, H.: Le concept d'anaphore, de cataphore et de déixis en linguistique française. Revue québécoise de linguistique **29**(2), 55–77 (2001)
13. Peroni, S., Shotton, D.: Fabio and cito: ontologies for describing bibliographic resources and citations. Web Semantics: Science, Services and Agents on the World Wide Web **17**, 33–43 (2012)
14. Ritchie, A., Robertson, S., Teufel, S.: Comparing citation contexts for information retrieval. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 213–222. ACM (2008)
15. Ritchie, A., Teufel, S., Robertson, S.: How to find better index terms through citations. In: Proceedings of the workshop on how can computational linguistics improve information retrieval? pp. 25–32. Association for Computational Linguistics (2006)
16. Singh, M., Patidar, V., Kumar, S., Chakraborty, T., Mukherjee, A., Goyal, P.: The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1271–1280. ACM (2015)
17. Teufel, S.: Argumentative zoning for improved citation indexing. In: Computing attitude and affect in text: Theory and Applications, pp. 159–169. Springer (2006)
18. Teufel, S., et al.: Argumentative zoning: Information extraction from scientific text. Ph.D. thesis, Citeseer (1999)