InTeReC: In-text Reference Corpus for Applying Natural Language Processing to Bibliometrics

Marc Bertin[1] and Iana Atanassova[2]

[1] ELICO Laboratory, Université Claude Bernard Lyon 1, France
marc.bertin@univ-lyon1.fr
[2] CRIT-Centre Tesnière, Université de Bourgogne Franche-Comté, France
iana.atanassova@univ-fcomte.fr

Abstract. Bibliometrics is more and more interested in the full text processing and the study of the structure of scientific papers. The contexts of in-text references present in articles are particularly relevant for such studies. This work describes the construction of the InTeReC dataset, which is an in-text reference corpus that aims to promote experimental reproducibility and to provide a standard dataset for further research. The InTeReC dataset is a set of sentences containing in-text references together with all the data necessary for their recontextualization in papers using standard CSV format. This should encourage the implementation of natural language processing tools for Bibliometric studies and related research in information retrieval and visualization.

Keywords: Bibliometrics, Citation Analysis, Citation Context Analysis, Information Analysis, Natural Language Processing, IMRaD

1 Introduction

The assumption that the contexts of the bibliographic references present in a scientific article play an important role in characterizing the relationship between citing works and cited works have been accepted for many decades. Publications are connected to each other by citations and citations contexts categorize the semantic relations that exist between them. Whether the study of citation contexts relies mainly on linguistic clues or machine learning techniques, citation contexts for each research experiment need to be extracted from scientific corpora. Also, the extraction of citation contexts is a preliminary step to any statistical, distributional, syntactic or semantic analysis.

Sentences containing in-text references may contain relevant information about the cited research and cited author's research areas. Recently, He and Chen [12] provides an approach to understanding citation contexts which characterizes the complex roles of a publication. If we are interested in the intellectual structure of a discipline, the analysis of co-citations has been widely studied. However, recent work highlights the interest of taking into account the full text and more precisely the paragraphs [13] of papers. Other approaches are interested in extracting information from publications and adding semantic attributes to in-text references that can be defined as traditional. For example, Parinov [17] focuses on papers' references, in-text references and citation contexts with the purpose to visualize citations relationships, their semantic attributes and related statistics as annotations.

In this paper, we propose a large scale dataset of citation contexts and explain the methods that were used for its construction. Other similar initiatives exist with various objectives. For example, the ESWC-14 Challenge: Semantic Publishing[3] – Assessing the Quality of Scientific Output (see [11]) focus on the extraction and assessment of workshop proceedings.The recent activity of research based on full text and the analysis of in-text references has lead to a race in the size of datasets. If we look at the size of the corpora used by the different actors of our community, we can see that the values are very heterogeneous. The corpus for the CL-SciSumm task[4] deal with automatic paper summarization in Computational Linguistics and is extracted from the ACL Anthology corpus and its citing papers [15]. In 2017, Hu et al. [14] worked with 350 articles from Journal of Informetrics. In 2013 Ding et al. [10] analyse 866 articles from JASIST. The largest study of in-text reference distributions in 2016 was proposed by Bertin et al. [7] who analysed 45,000 papers published in the PLOS journals. Recently, Boyack et al. [9] focused on the PubMed Central Open Access Subset and Elsevier journals with five million full text records for in-text reference analysis. It is clear that we observe an increase in the size of the textual data but also a methodological evolution in the processing capabilities, advocating for larger datasets and the use of statistical tools. In general, the construction of a corpus is a heavy task and requires skills and means that can be important.

In this paper we describe the creation of the InTeReC dataset [6], which is a corpus of sentences containing in-text references extraction from papers published by PLOS. This dataset is available at https://zenodo.org/record/1203737. Here we will not detail or define what a corpus is. For this we can refer for example to Lüdeling and Kytö (see [16]).

The aim of this paper is to present the method of the construction of the InTeReC corpus, which is a standard in-text reference corpus taking into account the different elements relevant to the implementation of experimental protocols. The overall objective is to facilitate citation context analyses and various distributional analyses by providing a large dataset to the community. The InTeReC dataset also serves the purposes of reproducibility, interoperability and cumulative research.

2 Method

The construction of the InTeReC dataset is based on several analyzes and experiments carried out in the recent years. In this section we summarize the methods that were used in order to propose a dataset that is reusable by the community for studying citation contexts.

[3] http://challenges.2014.eswc-conferences.org/index.php/SemPub

[4] http://wing.comp.nus.edu.sg/ cl-scisumm2018/

Working with the full text of papers, we first classify the section titles in order to identify the four major section types in the IMRaD sequence (Introduction, Methods, Results and Discussion). This categorization aims to verify the coherence of the corpus with the IMRaD structure. In many articles, the four section types exist but not in the same order. For the InTeReC dataset, we focused only on paper that follows the IMRaD structure, i.e. papers that contain the four section types in the correct order. We then process the text content of all paragraphs and segment them into sentences. In our approach, sentences are considered as the basic textual units and are used to express the positions of references in the article and in the section. This approach allows for example to assign relative positions of all references and to obtain the distribution of references along the text [7]. Finally, we count the number of references in each sentence. The InTeReC dataset contains only sentences are to have one single in-text reference.

The links between the in-text references and the cited papers or bibliography items are preserved throughout the processing.

2.1 Data: source and structure

For this corpus, we have used the entire set of research articles published by PLOS[5] up to September 2013. This initial corpus contains 90,071 articles.

As these 7 journals follow the same publication model but are in different scientific fields, our aim is to observe the different uses of bibliographic references in these fields and their relation to the structure of the articles.

PLOS provides access to the articles in the XML format. The set of XML elements and attributes that are used for the representation of journal articles are known as Journal Article Tag Suite (JATS), which is an application of Z39.96-2012. Technology evolves quickly and we have to take into consideration that JATS is a continuation of the NLM Archiving and Interchange DTD works by NCBI[6]. As this format is also used by PubMed, this work can easily be extended to processing the PubMed Open Access Subset which is a larger dataset. The JATS structure of an article consists of three main elements: *front – body – back*, and the textual content of the article is in the *body* element. It is further divided into sections and paragraphs. The *front* element contains some traditional metadata fields (title, authors, etc.) as well as the article type.

Different types of articles are present in the corpus, such as "Research article", "Synopsis", "Primer", "Essay", and the typology is given in the article's metadata. We have focused on the "Research article" type, obtaining a total of 85,660 articles out of the initial 90,071 articles in the corpus.

[5] Founded in 2001, the Public Library of Science (PLOS) is an Open Access publisher of seven peer-reviewed academic journals, mostly in the fields of Biology and Medicine. *PLOS ONE*, the publishers' general journal covers, however, all fields of science and social sciences.

[6] http://dtd.nlm.nih.gov

2.2 Segmentation and section title processing

One of our objectives is to identify the rhetorical structure of the articles. The use of the IMRaD sequence (Introduction, Methods, Results and Discussion) is part of the editorial requirements of the PLOS journals and the large majority of articles include these four sections. In some articles however, the sections are not always in the same order.

Sections are represented as separate elements in the original XML files. The research articles in the corpus contain a total of 404 311 sections. We categorized them automatically by analyzing the section titles in order to match the existing sections with one of the section types in the IMRaD structure [1,2]. In fact, variations can exist in the ways authors choose to title the sections, e.g. the Methods section can have titles such as "Materials and Methods", "Method and Model", etc. We have constructed a set of regular expressions in order to classify the sections automatically. Table 1 presents some basic statistics of the result of this classification. The last two classes, (MR) and (RD), appear in some articles where one section merges two of the main section types and thus the article contains only three main sections.

Class	Section type	Number of sections
I	Introduction	83,961
M	Methods	84,006
R	Results	76,909
D	Discussion	76,964
(MR)	Methods and Results	32
(RD)	Results and Discussion	7,072
Total		*328,944*

Table 1: Classification of section titles according to the four section types of IMRaD

We further restricted the set of sentences to be included in the InTeReC dataset by selecting only sentences that have a single citation and that contain at least one occurrence of the most frequent verbs that have been attested in citation contexts. These steps are explained in the following subsections.

Each paragraph was segmented into sentences by analyzing the punctuation of the text following a set of typographic rules. All the occurrences of symbols denoting sentence boundaries (point, exclamation mark, etc.) were examined and disambiguated. In fact, the occurrence of a point in a text does not necessarily mean a sentence end, because in many cases it can be part of an abbreviation, references, genus species, numeric values, etc. We used a set of finite-state automata in order to determine the contexts in which the points signal sentence ends.

2.3 Article structures

Once we classified the sections, we examined the sequence of sections present in each article. To produce the InTeReC dataset, we focused only on articles where the order of the four sections is: "I,M,R,D". Considering merged sections, there are three possible article structures, that are listed in table 2. The last two columns of this table give the total number of sentences in the articles and the number of sentences that contain at least one in-text reference.

Article structure	Articles	Sentences	Sentences with references
I,M,R,D	44,370	7,656,518	1,704,326
I,M,(RD)	2,971	504,246	113,237
I,(MR),D	28	5,300	937
Total	*47,369*	*8,166,064*	*1,818,500*

Table 2: Article structures following the IMRaD sequence

The following processing was done on these 47,369 articles from which was selected the InTeReC dataset.

2.4 Reference processing

Our algorithm examines each sentence and counts the number of references present in the text. In fact, the input data is in the XML format where the references are represented in <*xref*> tags. Our algorithm covers all possible typographic variations for reference ranges and infers the missing data from the input XML. As a result we obtain the list of sentences in the text, where to each sentence we have associated a reference count as well as a list of reference identifiers corresponding to the bibliography entries.

We note that counting the <*xref*> tags are not a reliable method to obtain the reference counts, especially if one is interested in multiple in-text references (MIR) [5]. When in-text references are in a numeric form, reference ranges are often present in sentences containing MIR. For example, in-text references such as *"[16]–[24]"* are represented by two *xref* elements that point to the corresponding bibliography items, while in fact there are 9 different citations, 7 of which are not present in the XML markup. In order to identify correctly MIR and their number in sentences it is important to detect in-text reference ranges.

For this first version of the InTeReC dataset we have chosen to include only sentences that contain one single reference. These citation contexts establish links between only two works, the cited work and the citing article, and thus we can consider them as the simplest cases to study in terms of citation context analysis.

2.5 Part-Of-Speech tagging and verb phrase extraction

A series of experiments have been published around verbs occurring in citation contexts and their distributions [3,4]. In general, verbs give important information about the nature of the relation between the article and the cited work. Polysemy is one possible problem when dealing with verbs, but in our case this phenomenon is reduced as we work specifically on citation contexts. Our hypothesis is that the semantic meaning of the relation that exists between the cited work and the citing article is often expressed, to some extent, by the verb phrase in the sentence containing the in-text reference. For this reason, we examined verb phrases that appear in the sentences that contain citations and included them in the dataset.

Bertin et al. [3,4]. published lists of most frequent verbs that appear in citation contexts with respect to section types in the IMRaD structure. We considered the most frequent verbs in each section that are given on table 3.

show	use	include
suggest	identify	find
require	associate	involve
lead	perform	follow
obtain	generate	base
determine	contain	calculate
carry	report	observe
express	see	

Table 3: Verbs that appear in citation contexts: most frequent verbs in the different sections of the IMRaD structure [3,4]

All sentences were processed using the Part-Of-Speech tagger of python NLTK[7] [8]. In the output verb forms are tagged by labels such as VB, VBD, VBG, VBN, VBP, VBZ that stands for base form, past tense, present participle, etc. We then identified verb phrases by producing parse trees using a simple grammar.

The InTeReC dataset contains sentences that contain occurrences of the verbs in table 3 and their verb phrases have been identified. By keeping only sentences that contain these verbs, we eliminate many sentences that contain perfunctory citations because they only mention the cited work without explicitely identifying the its relation with the article. Thus we obtained the final set of 314,023 sentences for the dataset.

3 InTeReC dataset structure

The InTeReC dataset contains a list of sentences in full text. Information is given on the position of the sentences relative to the article and the section

[7] https://www.nltk.org

in which they appear, the section type with respect to the four main types of the IMRaD structure, as well as verb phrases that occur in the sentence. Each sentence contains one single in-text reference.

The dataset is published in the CSV format [6], with the following column list:

journal: journal title
doi: DOI of the article from which the sentence was extracted
article-length: size of the article, as number of sentences
article-pos: position of the sentence in the article, as number of sentences from the beginning of the article
section-length: size of the section, as number of sentences
section-pos: position of the sentence in the section, as number of sentences from the beginning of the section
section-type: section type (one of: I, M, R, D, MR, RD)
sentence-text: full text of the sentence
verb-phrases: a list of verb phrases that occur in the sentence, comma separated

The format of the dataset has been chosen to facilitate the exploitation of the data and make it compatible and easily reusable for most types of processing.

4 Discussion and Conclusion

Although this corpus takes many aspects into account, it is not exhaustive and has characteristics that underline the inherent limitations of this approach. For example, we have limited this work to level of the sentence, and thus do not take into account relevant citation contexts that span across sentence boundaries through the use of anaphora.

Many improvements are possible and they are currently receiving our attention in the evolution of our corpus. The first concerns the quantitative aspect with the extension of the corpus to new sources. For this, we can take into account for example PubMed[8], arXiv[9] and the CEUR Workshop Proceedings[10].

The second evolution concerns a more qualitative aspect of the dataset and adding semantic annotations. The main idea is to implement semantic annotation in order to provide an training dataset for supervised learning tools. For this purpose, we investigate tools to annotate segments with precision values close to those of human annotators. Indeed, one challenge would be proposing semantic annotations for ontologies such as CiTO (see [18]).

Finally, the InTeReC dataset can be accessed and visualized using R-shiny[11] interface in order to provide users a way to interact with the data and observe the distributional phenomena.

[8] www.nlm.nih.gov/databases/download/pubmed_medline.html
[9] https://arxiv.org
[10] http://ceur-ws.org
[11] https://shiny.rstudio.com

This dataset aims to facilitate the reproducibility of future research on in-text citation analysis and thus provide a common foundation for the development of a unified model of citation context analysis.

5 Acknowledgments

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

References

1. Atanassova, I., Bertin, M.: Semantic Web Evaluation Challenge: SemWebEval 2014 at ESWC 2014, chap. Semantic Facets for Scientific Information Retrieval, pp. 108–113. Communications in Computer and Information Science (Book 475), Springer, Anissaras, Crete, Greece (May 25-29 2014)
2. Bertin, M., Atanassova, I.: Semantic Web Evaluation Challenge: SemWebEval 2014 at ESWC 2014, chap. Extraction and Characterization of Citations in Scientific Papers, pp. 120–128. Communications in Computer and Information Science (Book 475), Springer, Anissaras, Crete, Greece (May 25-29 2014)
3. Bertin, M., Atanassova, I.: A study of lexical distribution in citation contexts through the imrad standard. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014). vol. 1143, pp. 5–12. CEUR Workshop Proceedings, Amsterdam, The Netherlands (April 13 2014)
4. Bertin, M., Atanassova, I.: Factorial correspondence analysis applied to citation contexts. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 37th European Conference on Information Retrieval (ECIR 2015). Vienna, Austria (March 29 2015)
5. Bertin, M., Atanassova, I.: Multiple in-text reference aggregation phenomenon. In: Proceedings of the 3rd Workshop on Bibliometric-enhanced Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016). pp. 14–22. Padua, Italy (2016)
6. Bertin, M., Atanassova, I.: InTeReC: In-text Reference Corpus - Single References Dataset (Mar 2018), https://doi.org/10.5281/zenodo.1203737
7. 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 (2016), http://dx.doi.org/10.1002/asi.23367
8. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
9. 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), http://www.sciencedirect.com/science/article/pii/S1751157717303516
10. Ding, Y., Liu, X., Guo, C., Cronin, B.: The distribution of references across texts: Some implications for citation analysis. Journal of Informetrics 7(3), 583 – 592 (2013), http://www.sciencedirect.com/science/article/pii/S1751157713000230

11. Dragoni, M., Solanki, M., Blomqvist, E.: Semantic Web Challenges: 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28-June 1, 2017, Revised Selected Papers, vol. 769. Springer (2017)
12. He, J., Chen, C.: Understanding the changing roles of scientific publications via citation embeddings. arXiv preprint arXiv:1711.05822 (2017)
13. Hsiao, T.M., Chen, K.h.: Yet another method for author co-citation analysis: A new approach based on paragraph similarity. Proceedings of the Association for Information Science and Technology 54(1), 170–178 (2017), http://dx.doi.org/10.1002/pra2.2017.14505401019
14. Hu, Z., Lin, G., Sun, T., Hou, H.: Understanding multiply mentioned references. Journal of Informetrics 11(4), 948–958 (2017)
15. Jaidka, K., Chandrasekaran, M.K., Rustagi, S., Kan, M.Y.: Overview of the cl-scisumm 2016 shared task. In: In Proceedings of Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries (BIRNDL 2016) (2016)
16. Lüdeling, A., Kytö, M.: Corpus linguistics: An international handbook. Citeseer (2008)
17. Parinov, S.: Semantic attributes for citation relationships: Creation and visualization. In: Garoufallou, E., Virkus, S., Siatri, R., Koutsomiha, D. (eds.) Metadata and Semantic Research. pp. 286–299. Springer International Publishing, Cham (2017)
18. Shotton, D.: CiTO, the citation typing ontology. Journal of biomedical semantics 1(Suppl 1), S6 (2010)