InTeReC: In-text Reference Corpus for Applying NLP to Bibliometrics

Marc Bertin & Iana Atanassova

ELICO Laboratory, Université Claude Bernard Lyon 1, France
CRIT-Centre Tesnière, Université de Bourgogne Franche-Comté, France

BIR-2018

Grenoble, March 26, 2018

Plan

Research problem

- 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;
- Publications are connected to each other by citations and citations contexts categorize the semantic relations that exist between them.

Objectives

- Propose a easy to use standard dataset of citation contexts, reusable for further research;
- Facilitate experimental reproducibility
- Encourage the implementation of Natural Language Processing tools for Bibliometric studies and related research in information retrieval and visualization.

Other existing resources

- **Corpora**: CL-SciSumm from the ACL Anthology corpus
- **Challenges**: ESWC-14 Challenge: Semantic Publishing – Assessing the Quality of Scientific Output
- **Researchers' corpora**:
  - Hu et al. (2017): 350 articles from Journal of Informetrics;
  - Ding et al. (2013): 866 articles from JASIST;
  - Boyack et al. (2018): 5M articles from PubMed Central Open Access Subset and Elsevier journals;
  - ...

Dataset source

- Seven peer-reviewed academic journals published in Open Access by the Public Library of Science (PLOS).
- We processed the entire dataset up to September 2013:
  - 90,071 articles
  - about 85,600 articles of type ”Research article”.
- Format: XML Journal Article Tag Suite (JATS)

Dataset statistics

Journal	Research articles	In-text references	Citation contexts
PLOS Biology	1,754	5,798,761	91,117
PLOS Comp. Bio.	2,560	7,894,013	126,870
PLOS Genetics	3,414	11,935,753	185,537
PLOS Medicine	926	2,060,487	34,819
PLOS Negl. Trop. Dis.	1,872	3,798,743	73,211
PLOS ONE	72,123	154,500,905	2,854,082
PLOS Pathogens	2,976	10,459,231	162,878
Total	85,625	196,447,893	3,528,514

Example


```
−<body>
  −<sec id="s1">
     <title>Introduction</title>
   −<p>
      In eukaryotes, DNA replication is initiated at multiple origins. Potential sites in the genome of the yeast
      <italic>Saccharomyces cerevisiae</italic>
      that may serve this function are referred to as autonomously replicating sequences, or ARS elements [
      <xref ref-type="bibr" rid="pcbi-0010007-b01">1</xref>
      ]. ARS elements are more A+T-rich than the genomic average, and contain regions of low local thermodynamic stability that are thought to be necessary for function [
      <xref ref-type="bibr" rid="pcbi-0010007-b02">2</xref>
      ,
      <xref ref-type="bibr" rid="pcbi-0010007-b03">3</xref>
      ]. However, the duplex unwinding required for replication initiation occurs as an isothermal process within topologically constrained domains of DNA. Under these conditions susceptibility to strand opening is not dependent only on local thermodynamic stability. Instead, superhelical stresses couple together the strand-opening behaviors of all base pairs that experience them. We hypothesize that the superhelical stresses that occur in vivo play a role in regulating the strand opening needed to initiate replication. This suggests that ARS elements should have an increased local susceptibility to superhelically induced duplex destabilization (SIDD). Here we demonstrate that virtually all known ARS elements do indeed show a significant local increase in
```

Segmentation

Sentence segmentation of all sections is carried out with two objectives:

- Identify sentences that are citation contexts;
- Calculate the positions of in-text references in the article and in the section, as number of sentences from the beginning.

Classification of sections

- Objective: identify the four main section types of the IMRaD sequence.
- Method: analyze section titles and use rules based on regular expressions to capture the possible variations in titles: e.g. ”Materials and Methods”, ”Method and Model”, ...
- Total number of sections: 404,311.
- Classified sections: 328,944.

Classification of sections: results

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

Article structures

Articles that follow the IMRaD sequence, in the same order:

Article structure	Articles	Sentences	Sent. 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*

Processing of in-text references

- In-text references are in represented as *xref* elements in the XML structure.
- We count the *xref* elements in sentences, and select sentences that have only 1 in-text reference.
- This method is not sufficient to process multiple in-text references (see Bertin et al 2016, BIR).

```
e.g.
"A number of recent studies have used a modification of the picture viewing
procedure by substituting pleasant pictures with photographs of loved, familiar
faces <xref ref-type="bibr" rid="pone.0041631-Bartels1">[16]</xref> –
<xref ref-type="bibr" rid="pone.0041631-Xu2">[24]</xref>."
```

Verbs in citation contexts

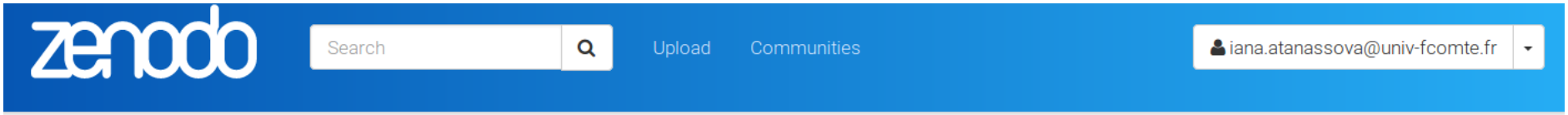
- Verbs give important information about the nature of the relation between the article and the cited work.
- Polysemy is one possible problem, but in our case it is reduced as we work specifically on citation contexts.
- Most frequent verbs in citation contexts (Bertin et al 2015, BIR):

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

Verb phrases selection

- We selected citation contexts that contain forms of the most frequent verbs.
- This step allows to eliminate some perfunctory citations: sentences that only mention the cited work without explicitly identifying its relation with the article.
- Sentences were processed using the Part-Of-Speech tagger of python NLTK, and verb phrases were identified by producing parse trees using a grammar.
- Final set of 314,023 sentences for the dataset.

Dataset



March 19, 2018

Dataset Open Access

InTeReC: In-text Reference Corpus - Single References Dataset

Bertin, Marc; Atanassova, Iana

This dataset contains a set of sentences extracted from articles published by the Public Library of Science (PLOS) up to September 2013. Information is given on the position of the sentences relative to the article and the section 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 in the CSV format. Size: 314023 sentences.

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 (see below)
- *sentence-text*: full text of the sentence
- *verb-phrases*: a list of verb phrases that occur in the sentence, comma separated

Possible section types are:

- I: Introduction

Indexed in OpenAIRE

Publication date:
March 19, 2018

DOI:
DOI 10.5281/zenodo.1203737

Keyword(s):
In Text References, Bibliometrics, Citation Analysis, IMRaD, Natural Language Processing, PLOS, Citation Context Analysis

Meeting:
Bibliometric-enhanced Information Retrieval: 7th International BIR workshop (7th BIR workshop) (BIR), Grenoble, France, 26 March 2018

License (for files):
Creative Commons Attribution-NonCommercial 4.0

https://zenodo.org/record/1203737, DOI: 10.5281/zenodo.1203737

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 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 CSV format, UTF-8.
- 314,023 sentences, ∼84MB.

InTeReC dataset structure: 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

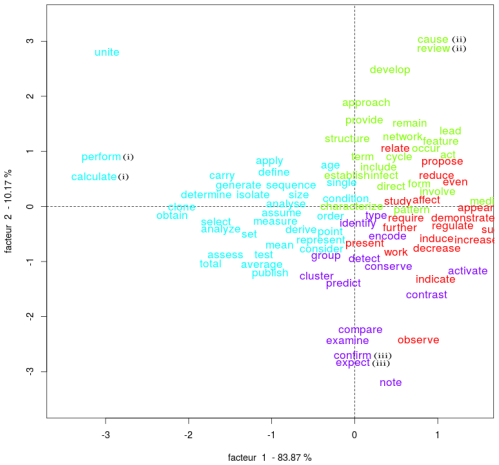
section-pos: position of the sentence in the section, as number of sentences from the beginning of the section

section-type: section type (one of: I, M, R, D, MR, RD)

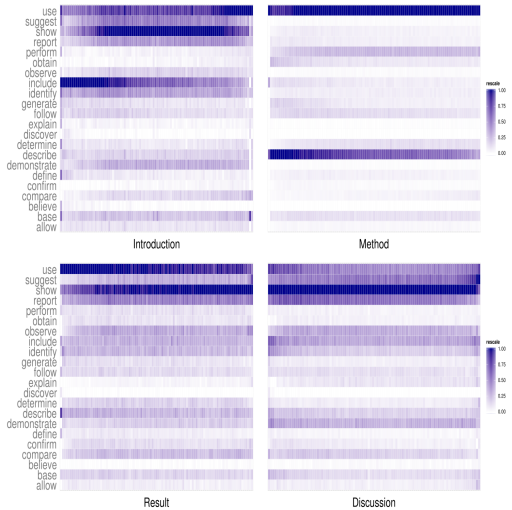
sentence-text: full text of the sentence

verb-phrases: a list of verb phrases that occur in the sentence, comma separated

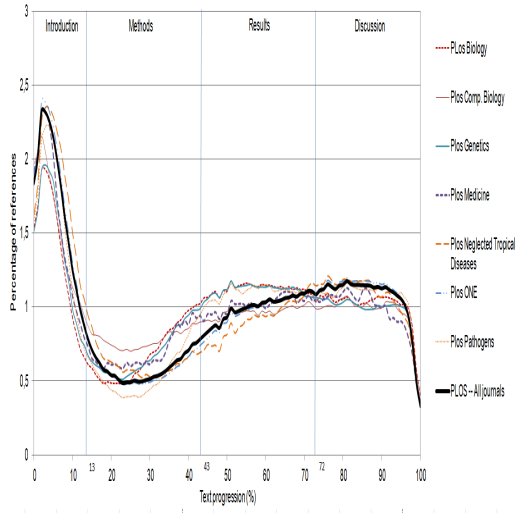
Some reproducible results:



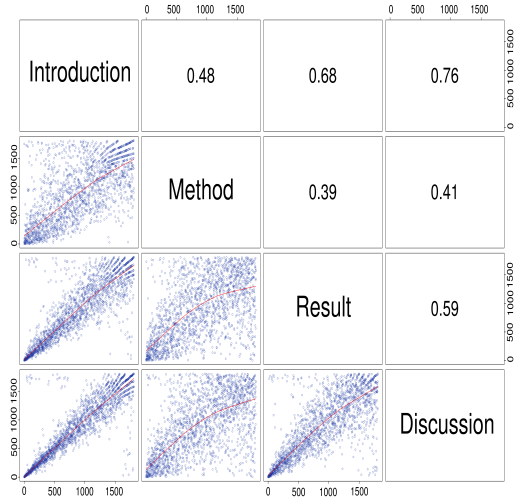
Factorial correspondence analysis applied to citation contexts [3]



A study of lexical distribution in citation contexts through the IMRaD standard [2]



The invariant distribution of references in scientific papers [6]



A study of lexical distribution in citation contexts through the IMRaD standard [2]

Perspectives

Enriching the dataset with:

- other article structures (e.g. R,I,M,D)
- multiple in-text references, ranges, etc., preserving links to cited papers
- DOI for cited papers
- ORCID for researcher
- semantic annotation, stored as RDF/OWL, queries with SparQL
- larger data sources, e.g. PubMed OA Subset, arXiv, ...

Thank you for your attention!

Marc Bertin & Iana Atanassova

marc.bertin@univ-lyon1.fr
iana.atanassova@univ-fcomte.fr

Related work

Iana Atanassova and Marc.

Temporal properties of recurring in-text references.
D-lib Magazine, 22(9/10), September/October 2016.

Marc Bertin and Iana Atanassova.

A study of lexical distribution in citation contexts through the IMRaD standard.
In *BIR*, pages 5–12, Amsterdam, The Netherlands, April 13 2014.

Marc Bertin and Iana Atanassova.

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.

Marc Bertin and Iana Atanassova.

The context of multiple in-text references and their signification.
International Journal on Digital Libraries, pages 1–12, 2017.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.

Mapping the Linguistic Context of Citations.
ASIS&T Bulletin, 41(2), December/January 2015.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.

The invariant distribution of references in scientific papers.
JASIST, 67(1):164–177, January 2016.

Marc Bertin, Iana Atanassova, Cassidy R. Sugimoto, and Vincent Larivière.

The linguistic patterns and rhetorical structure of citation context: an approach using n-grams.
Scientometrics, (109):1417–1434, September 2016.