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

Grenoble, March 26, 2018
Plan

1. Objectives
2. Dataset source
3. Methods and processing steps
4. InTeReC Dataset
5. Perspectives
Objectives

Dataset source

Methods and processing steps

InTeReC Dataset

Perspectives
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;
  - ...

Marc Bertin & Iana Atanassova
InTeReC: In-text Reference Corpus
BIR 2018 - Grenoble
1 Objectives

2 Dataset source

3 Methods and processing steps

4 InTeReC Dataset

5 Perspectives
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

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

2 Dataset source

3 Methods and processing steps

4 InTeReC Dataset

5 Perspectives
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

<table>
<thead>
<tr>
<th>Class</th>
<th>Section type</th>
<th>Number of sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Introduction</td>
<td>83,961</td>
</tr>
<tr>
<td>M</td>
<td>Methods</td>
<td>84,006</td>
</tr>
<tr>
<td>MR</td>
<td>Methods and Results</td>
<td>32</td>
</tr>
<tr>
<td>R</td>
<td>Results</td>
<td>76,909</td>
</tr>
<tr>
<td>RD</td>
<td>Results and Discussion</td>
<td>7,072</td>
</tr>
<tr>
<td>D</td>
<td>Discussion</td>
<td>76,964</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>328,944</td>
</tr>
</tbody>
</table>
### Article structures

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

<table>
<thead>
<tr>
<th>Article structure</th>
<th>Articles</th>
<th>Sentences</th>
<th>Sent. with references</th>
</tr>
</thead>
<tbody>
<tr>
<td>I, M, R, D</td>
<td>44,370</td>
<td>7,656,518</td>
<td>1,704,326</td>
</tr>
<tr>
<td>I, M, (RD)</td>
<td>2,971</td>
<td>504,246</td>
<td>113,237</td>
</tr>
<tr>
<td>I, (MR), D</td>
<td>28</td>
<td>5,300</td>
<td>937</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>47,369</strong></td>
<td><strong>8,166,064</strong></td>
<td><strong>1,818,500</strong></td>
</tr>
</tbody>
</table>
Methods and processing steps

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-Bartels">[16]</xref> -- <xref ref-type="bibr" rid="pone.0041631-Xu">[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.
Objectives

Dataset source

Methods and processing steps

InTeReC Dataset

Perspectives
InTeReC: In-text Reference Corpus - Single References Dataset

Marc Bertin, Marc; Iana 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
- **dof**: 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

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
Objectives

Dataset source

Methods and processing steps

InTeReC Dataset

Perspectives
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]
Enriching the dataset with:

- other article structures (e.g. R,l,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.

Marc Bertin and Iana Atanassova.
Factorial correspondence analysis applied to citation contexts.

Marc Bertin and Iana Atanassova.
The context of multiple in-text references and their signification.

Marc Bertin, Iana Atanassova, Vincent Larivi`ere, and Yves Gingras.
Mapping the Linguistic Context of Citations.

Marc Bertin, Iana Atanassova, Vincent Larivi`ere, and Yves Gingras.
The invariant distribution of references in scientific papers.

Marc Bertin, Iana Atanassova, Cassidy R. Sugimoto, and Vincent Larivi`ere.
The linguistic patterns and rhetorical structure of citation context: an approach using n-grams.