Hybrid Approach for the Semantic Processing of Scientific Papers

Marc Bertin and Iana Atanassova

CIRST, Université du Québec à Montréal
B.P. 8888, Succ. Centre-ville
H3C 3P8 Montreal (Quebec) Canada
bertin.marc@courrier.uqam.ca, iana.atanassova@nlp-labs.org

Abstract. We propose a hybrid method for the extraction and characterization of citations in scientific papers using machine learning combined with rule-based approaches. Our protocol consists of the extraction of metadata, bibliography parsing, section titles processing, and find-grained semantic annotation on the sentece level of texts. This allows us to generate Linked Open Data from a set of research papers in XML.

Keywords: Semantic Annotation, Citation Acts, CRF, RDF graphs

1 Introduction

With the advent of Open Science and the free access to standardized scientific articles, it becomes possible to explore and process massive amounts of textual data. Many studies deal with the text mining and semantic annotation of scientific papers. The Task 2 of ESWC-14 Semantic Publishing Challenge focuses on the extraction of information about in-text citations and bibliographic references in scientific articles and their relevance. Several different types of processing are necessary: citation extraction and characterization, analysis of author names, identification of grants and funding agencies, identification of literary review sections.

2 Method

The protocol described below is based on the XML parsing, segmentation and semantic annotation of the text. We have developed a stand-alone application in Java for the processing of scientific papers that uses several other software libraries available in open source. Figure 1 gives an overview of the processing stages that will be detailed below.

Dataset. Two datasets were provided for the task: a Training Dataset of 150 papers from 15 different journals and an Evaluation Dataset which is a superset of 400 papers from 71 journals. The documents are in XML JATS (Journal

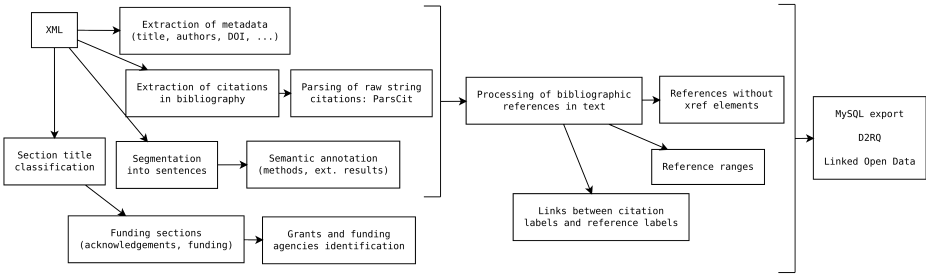
XML
Extraction of metadata (title, authors, DOI, ...)
Extraction of citations in bibliography
Parsing of raw string citations: ParsCit
Section title classification
Segmentation into sentences
Semantic annotation (methods, ext. results)
Funding sections (acknowledgements, funding)
Grants and funding agencies identification
Processing of bibliographic references in text
References without xref elements
Reference ranges
Links between citation labels and reference labels
MySQL export
D2RQ
Linked Open Data

Fig. 1. Processing stages

Article Tag Suite) and TaxPub, an official extension of JATS customized for taxonomic treatments.

Our processing relies on the full segmentation into sentences of all paragraphs. Thus we use sentences as the basic unit for the further processing stages. The segmentation method consists of an analysis of the punctuation and capitalization patterns in texts. This approach has already been used to process the PLOS journals (see [1]). Table 1 presents the number of articles processed for Task 2, as well as the average number of sentences, citations and references in the bibliography per article.

Table 1. Datasets

Dataset	Articles	Avg sentences	Avg in-text citations	Avg references
Training	150	202.64	69.39	36.13
Evaluation	400	170.23	63.12	41.56

Extraction of metadata and bibliography parsing. Each article's metadata are extracted from the `front` element in the XML document. Paper identifiers, namely DOI and PubMedId are also extracted. If an identifier is missing, we try to recover it using the PubMed ID Converter API[1].

The processing of the bibliography items is more complex. In several journals reference elements were present as raw text strings in the bibliography. In such cases, in order to identify the metadata for each bibliography item, we use ParsCit (see [2,3]), which is an open source CRF reference parser. As ParsCit processed successfully only about 30% of these references, we have designed another algorithm based on rules for the analysis of the punctuation in the strings. We identify first the year in the reference string and then analyze the substrings to the left and to the right of the year. Our algorithm identifies author names, year and titles successfully for more than 60% of the references. However, it fails for

[1] PubMed API: `http://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/`

some of the references processed by ParsCit, which means that the combination of the two approaches gives better results than any of the two algorithms separately. Table 2 presents the number of successfully parsed references by ParsCit, by our algorithm and the remaining references that could not be parsed.

Table 2. Number of parsed raw string references using ParsCit

Dataset	Total references	Raw string references	Parsed by ParsCit	Parsed by our algorithm	Not parsed
Training	10,408	104	37 (35.58%)	85 (81.73%)	9 (8.65%)
Evaluation	25,246	471	136 (28.87%)	292 (62.00%)	134 (28.45%)

Processing of citations in text. The XML schema provides `xref` elements for citations that appear in body paragraphs. Each `xref` element contains an attribute that points to an item in the bibliography allowing to link the reference with the corresponding citation. However, two different problems can occur.

Firstly, some citations in texts are not identified as `xref` elements in the corpus. To resolve this, we have used regular expressions in order to identify strings that are likely to be citations in the text. For further disambiguation, the obtained strings were then matched against reference labels, first author names and years in the bibliography. Thus we were able to establish the links between the new citations and the reference.

Secondly, multiple citations can be grouped in a range (e.g. *"[10-13]"*) and in such cases, depending to the journal, pointers to all cited papers may or may not be present in the `xref` elements. We have analyzed the punctuation patterns between each two `xref` elements in the text, and identified possible citation ranges. All citations in the ranges were then matched against reference labels in the bibliography to establish the new links. Table 3 presents the number of new citations and links found by our processing methods.

Table 3. New citations and links found in the corpus

Dataset	New citations	New links between references and citations
Training	52	697
Evaluation	107	2,071

Section titles. The categorization of section titles is useful for identifying the literary review sections and also for retrieving the grants and funding agencies that are given either in the acknowledgments or in a separate funding section. We have categorized the section titles into 18 categories such as *introduction, method, analysis, funding.*

Bertin et al. [1] have shown that the distribution of reference citations in scientific papers is closely related to the cognitive structure. Their results point out that the introductory sections, where we could expect to find the literary review, tend to have significantly higher concentration of citations than the rest of the papers. However, not every article contains a literary review section and in general the section with the highest density of citations is not guaranteed to be the literary review section. For this reason, to identify literary review sections we consider two criteria. Firstly, the average number of citations per sentence in the section must be 1 or higher. Secondly, the section title must be categorized as *introduction/background* or *related literature*. About 22% of the sections could not be categorized and for those we rely only on the first criterion. If more than one sections correspond to these criteria, we take the section with the higher density of citations.

Grants and funding agencies processing. We suppose that information about grants and funding agencies is found in the acknowledgments or in the *Funding* section. To extract grant identifiers, we first filter the sentences that are likely to contain grants. We use linguistic clues, implemented by regular expressions, such as "*(was|is|were) (supported|funded) by*". In this way, we limit the processing to only a small part of the sentences and reduce the noise. The next steps are identifying the grants and the organizations in the sentences.

As grant identifiers are literals containing numbers, letters and dashes, they are recognized using a set of regular expressions. Then, to identify funding agencies we used Stanford Named Entity Recognizer[2] (see [4]) that allowed us to identify organization names.

Semantic annotation of methods and extended results. In order to characterize the function of citations, we need to take into consideration the semantic relations in their contexts. To do this, we use a rule-based semantic annotation approach [5] which relies on an linguistic ontology of citation acts. By using the previous work of Bertin et al. [6], who characterizes citations related to methods, we have extended this approach to process the extending results of a paper.

SPARQL queries and web interface. The data obtained by our processing is exported to a MySQL relational database and as RDF graph [7,8] using D2RQ[3].

The natural language queries for the task were modeled with SPARQL. The web interface that shows the SPARQL queries and their output is accessible on: `http://sempub2014.nlp-labs.org/task2/`, optimized for Mozilla Firefox.

3 Results and Discussion

Our implementation uses a hybrid approach combining rule-based and machine learning approaches. The quality and consistency of the input data, especially

[2] Stanford NER: `http://nlp.stanford.edu/ner/`

[3] D2RQ: `http://d2rq.org/`

the correct identification of citations and reference metadata in the papers, is crucial to the output. For this reason, the most important steps is our processing are the initial parsing and citation identification steps. The choice to work on the sentence level rather than on the paragraph level of texts allows to develop more fine-grained annotation and in some cases limit the noise. It also opens the possibility for other applications such as sentence extraction and document syntheses.

In producing our RDF data we use our proper ontology that is specifically designed to cover all of the task requirements. However, a mapping to already existing ontologies could be done for better interoperability with other tools. For example, we could consider CiTO [9] for the characterization of citations, DoCO [10] for document components and BiRO (`http://purl.org/spar/biro`) for the description of bibliographic references.

Some of the processing steps that we describe can be improved by machine learning from datasets specifically designed for these tasks. For example, given the similarity in the bibliographic patterns throughout the corpus, the parser of raw string references could be improved by using the set of references that are already parsed in the corpus as training dataset.

References

1. Bertin, M., Atanassova, I., Lariviere, V., Gingras, Y.: The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. In: Proceedings of the 14th ISSI Conference. (2013) 591–603
2. Councill, I.G., Giles, C.L., Kan, M.Y.: ParsCit: an Open-source CRF Reference String Parsing Package. In: LREC. (2008)
3. Do, H.H.N., Chandrasekaran, M.K., Cho, P.S., Kan, M.Y.: Extracting and Matching Authors and Affiliations in Scholarly Documents. In: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, ACM (2013) 219–228
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2005) 363–370
5. Desclés, J.P.: Contextual Exploration Processing for Discourse and Automatic Annotations of Texts. In: FLAIRS Conference. (2006) 281–284
6. Bertin, M., Atanassova, I., Descles, J.P.: Automatic Analysis of Author Judgment in Scientific Articles based on Semantic Annotation. In: 22nd International Florida Artificial Intelligence, Research Society Conference, Sanibel Island, Florida, AAAI Press (2009)
7. Bizer, C.: The D2RQ Platform-Treating Non-RDF Databases as Virtual RDF Graphs. www4.wiwiss.fu-berlin.de/bizer/d2rq (2004)
8. Cyganiak, R.: The D2RQ Platform-Accessing Relational Databases as Virtual RDF Graphs 2012. (2012)
9. Shotton, D.: CiTO, the Citation Typing Ontology. Journal of biomedical semantics **1**(Suppl 1) (2010) S6
10. Shotton, D., Peroni, S.: DoCO, the Document Components Ontology. (2011)