Identifying functions of citations with CiTalO

Angelo Di Iorio[1], Andrea Giovanni Nuzzolese[1,2], and Silvio Peroni[1,2]

[1] Department of Computer Science and Engineering, University of Bologna (Italy)
[2] STLab-ISTC Consiglio Nazionale delle Ricerche (Italy)
diiorio@cs.unibo.it, nuzzoles@cs.unibo.it, essepuntato@cs.unibo.it

Abstract. Bibliographic citation is one of the most important activities of an author in the production of any scientific work. The reasons that an author cites other publications are varied: to gain assistance of some sort, to review, critique or refute previous works, etc. In this paper we propose a tool, called *CiTalO*, to infer automatically the nature of citations by means of Semantic Web technologies and NLP techniques. Such a characterisation makes citations more effective for linking, disseminating, exploring and evaluating research.

1 Introduction

Bibliographic citations are the most used tools of academic communities for *linking* research, for instance by connecting scientific papers to related works or sources of experimental data. Citations are also tools for *disseminating*, as largely discussed in [9], and *exploring* research, for instance providing new interfaces for browsing data. Finally, citations are useful for *evaluating* research, e.g. through bibliometric measures such as *h-index* and *impact factor*.

All these activities can be radically improved by exploiting the actual "nature" of citations, i.e. the "author's reason for citing a given paper" [11]. The mere existence of a citation, in fact, does not provide any information about the reasons the author had in mind when creating that citation to some particular document rather than to another. It is the characterization of a citation that really capture its meaning and effect.

The goal of this paper is to present *CiTalO*, a tool that automatically annotates citations with properties defined in *CiTO* (*Citation Typing Ontology*)[3] [7]. These properties describe the nature of citations in scholarly works.

CiTalO is implemented in Java and can be used as either stand-alone component or web service. A demo version is also available at `http://wit.istc.cnr.it:8080/tools/citalo`: users can use a simple HTML form to submit an English sentence containing a citation to CiTalO and to receive the list of *CiTO* properties that characterize the nature of that citation. Multiple configurations can also be tested by using the same prototype. CiTalO exploits Semantic Web technologies and NLP techniques to produce the output. The tool is designed as a *chain of analysers* that (i) produce ontological statements from texts, (ii) search

[3] CiTO: http://purl.org/spar/cito.

patterns in those statements, (iii) maps those patterns into linguistic resources and (iv) use these resources to produce the final characterization conform to CiTO. The chain also includes a sentiment-analysis module to refine results.

The paper is structured as follows. In Section 2 we introduce previous works on classification of citations. In Section 3 we describe CiTalO introducing its structure. In Section 4, we conclude the paper sketching out some future works.

2 Related works

In [3] Copestake *et al.* introduce the SciBorg framework, which includes a module for discourse and citation analysis that follows the *Argumentative Zoning* scheme proposed by Teufel *et al.* [10] and produces quite good results.

Teufel *et al.* present a study about *function* of citations [11]. They provide a categorisation of possible citation functions organised in twelve classes, in turn clustered in *Negative*, *Neutral* and *Positive* rhetorical functions. They also performed some tests on hundreds of articles in computational linguistics, evaluating the output of several human annotators and a novel machine learning approach, and showed that the agreement between humans is actually higher than the agreement between humans and automatic analysis. Along the lines of the latter work, also Jorg analysed several documents within the ACL Anthology Networks[4] with the intent of identifying verbs usually used to carry important information about the nature of citations [6].

Closely related to the annotation of citation functions, in [2] Athar *et al.* propose and evaluate (with good result) a sentiment-analysis approach to citations, so as to identify whether a particular act of citing was done with positive (e.g. praising a previous work on a certain topic) or negative intentions (e.g. criticising the results obtained through a particular method).

3 CiTalO

CiTalO tries to guess the function of citations by combining techniques of ontology learning from natural language, sentiment-analysis, word-sense disambiguation, and ontology mapping. These techniques are thought to be applied in a pipeline whose input is the sentence of an article containing the citation – e.g. "It extends the research outlined in earlier work X", where *X* is a reference to a particular bibliographic entity – and the output is one or more properties of the CiTO ontology [7] – *cito:extends* for the previous example. The overall architecture is shown in Fig. 1, while an extensive explanation of features and drawbacks of CiTalO can be found in [4].

Sentiment-analysis for gathering the polarity of citation functions. The aim of this step is to capture the sentiment polarity emerging from the text in which the citation is included. This is connected to the classification of CiTO properties provided in [7], where the semantics of rhetorical citations is expressed

[4] ACL Anthology Network: http://clair.eecs.umich.edu/aan/index.php.

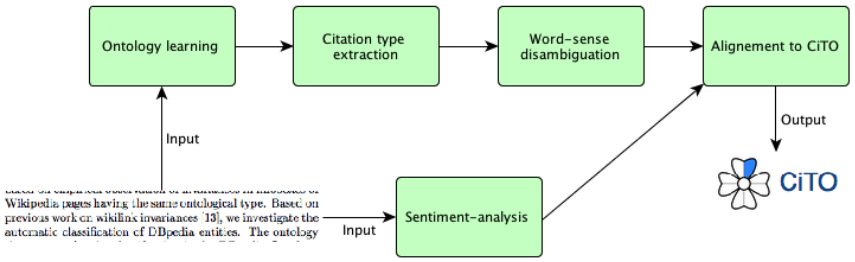

Fig. 1. The pipeline used by CiTalO. The input is the textual context in which the citation appears and the output is a set of properties of CiTO.

according to three different polarities, i.e. positive, neuter and negative. Being able to recognize the polarity behind the citation, in fact, would restrict the set of possible target properties from the CiTO ontology to match. Notice also that such an analysis goes in parallel with the others in CiTalO, being it a refinement filter of the results. The current sentiment-analysis component is based on AlchemyAPI[5] but it can be easily replaced with other similar tools.

Ontology extraction from the textual context of the citation. The first mandatory step of CiTalO consists of deriving a logical representation of the sentence containing the citation. This ontology extraction is performed by using FRED [8], a tool for ontology learning based on discourse representation theory, frames and ontology design patterns. The transformation of the sentence into a logical form allows us to recognize graph-patterns in order to detect possible types of rhetorical denotation of the citation. Consider, for instance, the sentence "it extends the research outlined in earlier work X", where "X" is the cited work. The graphical representation of the output in FRED, that is also available as RDF statements, is presented in Fig. 2.

Citation type extraction through pattern matching. The second step consists of extracting candidate types for the citation, by looking for patterns in the FRED result. We designed several graph-pattern-based heuristics by following similar criteria as lexico-syntactic patterns [1], extended with the exploitation of RDF graph topology and OWL semantics. These heuristics are implemented as SPARQL queries and some example are shown below:

```
SELECT ?type WHERE {?subj ?prop fred:X . ?subj a ?type}
SELECT ?type WHERE {?subj ?prop fred:X . ?subj a ?typeTmp .
  ?typeTmp rdfs:subClassOf+ ?type}
SELECT ?type WHERE {?subj a dul:Event . ?subj a ?type .
  FILTER(?type != dul:Event)}
SELECT ?type WHERE {?subj a dul:Event . ?subj a ?typeTmp .
  ?typeTmp rdfs:subClassOf+ ?type.FILTER(?type != dul:Event)}
SELECT ?type WHERE {?subj a dul:Event .
  ?subj boxer:patient ?patient . ?patient a ?type}
```

[5] AlchemyAPI: http://www.alchemyapi.com.

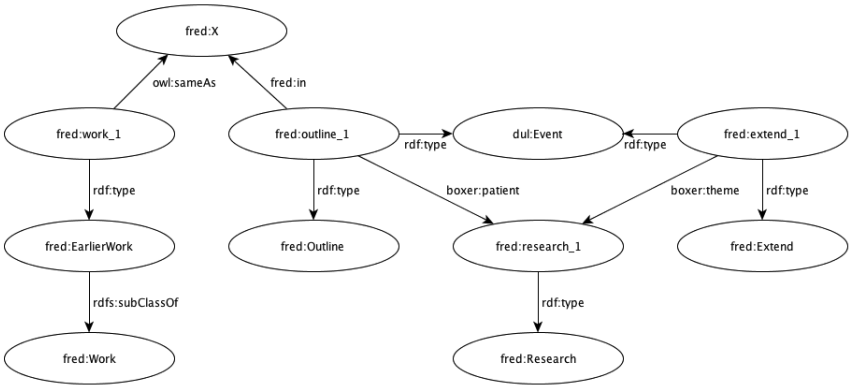

Fig. 2. FRED result for "It extends the research outlined in earlier work X".

The intended semantics of the above patterns is to select from the RDF graph all the types and their eventual taxonomies related to (i) the cited document, (ii) the events recognized into the citation, and the entities affected by those events (i.e. the entities playing the VerbNet role of being patient).

Applying these patterns to graph shown in Fig. 2, the following candidate types are found: *Outline, Extend, EarlierWork, Work, and Research.* The current set of heuristics is quite simple and incomplete, but we are continuously updating the catalogue by both investigating new heuristics.

Word-sense disambiguation. The next step consists of disambiguating the sense of each candidate type. This can be done through word-sense disambiguation services and APIs – in CiTalO we use IMS [12]. The disambiguation is performed with respect to OntoWordNet [5] and produces a list of synsets for the candidate types. Going back to the example, this phase would produce the following list[6]: (i) *Extend* is disambiguated as `own:synset-prolong-verb-1`, (ii) *Outline* as `own:synset-delineate-verb-3`, (iii) *Research* as `own:synset-research-noun-1`, (iv) *EarlierWork* and *Work* as `own:synset-work-noun-1`.

Alignment to CiTO. The last step consists of associating each synset to a CiTO property and refining results by using citation polarities and factual characterisation. We use two ontologies for this purpose: *CiTO2Wordnet* and *CiTOFunctions*. CiTO2Wordnet[7] maps all the CiTO properties defining citations with the appropriate Wordnet synsets [5]. CiTOFunctions[8] classifies each CiTO properties according to their factual and rhetorical functions [7]. The final alignment to CiTO is performed by means of a SPARQL CONSTRUCT query that uses the enhanced RDF graph obtained during the pipeline, the RDF graph of the polarity, OntoWordNet and the two ontologies just described.

[6] The prefix *own* stands for http://www.w3.org/2006/03/wn/wn30/instances/.
[7] CiTO2Wordnet ontology: http://www.essepuntato.it/2013/03/cito2wordnet.
[8] CiTOFunctions: http://www.essepuntato.it/2013/03/cito-functions.

4 Conclusions

CiTalO integrates Semantic Web technologies and NLP techniques to extract information about the nature, the motivations and the goals of each citation. The CiTalO architecture is composed of a pipeline of modules that map documents into ontological data, ontological data into linguistic resources and, finally, linguistic resources into CiTO properties.

The implementation is still at an early stage. On the other hand, the overall approach is very open to incremental refinements. We are currently working to improve *patterns' matching* phases in CiTalO and to include a mechanism for the automatic identification of *textual context* of citations given an input article. We also plan to perform exhaustive tests with a large set of documents and users.

References

1. Aguado de Cea, G., Gómez-Pérez, A., Montiel-Ponsoda, E., Suárez-Figueroa, M. C. (2008). Natural Language-Based Approach for Helping in the Reuse of Ontology Design Patterns. In Proceedings of EKAW 2008: 32-47. DOI: 10.1007/978-3-540-87696-0_6
2. Athar, A., Teufel, S. (2012). Context-Enhanced Citation Sentiment Detection. In Proceedings of HLT-NAACL 2012: 597-601.
3. Copestake, A., Corbett, P., Murray-Rust, P., Rupp, C. J., Siddharthan, A., Teufel, S., Waldron, B. (2006). An architecture for language processing for scientific text. In Proceedings of the UK e-Science All Hands Meeting 2006.
4. Di Iorio, A., Nuzzolese, A. G., Peroni, S. (2013). Towards the automatic identification of the nature of citations. To appear in Proceedings of SePublica 2013.
5. Gangemi, A., Navigli, R., Velardi, P. (2003). The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In Proceedings of CoopIS/DOA/ODBASE 2003: 820-838. DOI: 10.1007/978-3-540-39964-3_52
6. Jorg, B. (2008). Towards the Nature of Citations. In Poster Proceedings of FOIS 2008.
7. Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics, 17 (December 2012): 33-43. DOI: 10.1016/j.websem.2012.08.001
8. Presutti, V., Draicchio, F., Gangemi, A. (2012). Knowledge extraction based on discourse representation theory and linguistic frames. In Proceedings of EKAW 2012: 114-129. DOI: 10.1007/978-3-642-33876-2_12
9. Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. In Learned Publishing, 22 (2): 85-94. DOI: 10. 1087/2009202
10. Teufel, S., Carletta, J., Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In Proceedings of the 9th Conference of the EACL 1999: 110-117.
11. Teufel, S., Siddharthan, A., Tidhar, D. (2006). Automatic classification of citation function. In Proceedings of EMNLP 2006: 103-110.
12. Zhong, Z., Ng, H. T. (2010). It Makes Sense: A wide-coverage word sense disambiguation system for free text. In Proceedings of ACL 2010, System Demonstrations: 78-83.