Characterising citations in scholarly articles: an experiment

Paolo Ciancarini[1,2], Angelo Di Iorio[1], Andrea Giovanni Nuzzolese[1,2], Silvio Peroni[1,2], and Fabio Vitali[1]

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

Abstract. This work presents some experiments in letting humans annotate citations according to CiTO, an OWL ontology for describing the function of citations. We introduce a comparison of the performance of different users, and show strengths and difficulties that emerged when using that particular model to characterise citations of scholarly articles.

Keywords: CiTO, act of citing, citation function, ontology, scholarly articles, semantic publishing, user testing

1 Introduction

The mere existence of a citation might not be enough to capture the relevance of the cited work. For instance, some simple questions arise: is it correct to count negative and positive citations in the same way? Is it correct to give self-citations the same weight of others? Is it correct to give a survey the same weight of a seminal paper, by only counting the number of times it has been cited? A more effective characterisation of citations opens interesting perspectives that go beyond the quantitative evaluation of research products, as highlighted in [2].

To this end, the first issue to address is to identify a formal model for characterising the nature of citations in a precise way – i.e. a citation model. The citation model has to capture the citation functions, i.e. "the author's reasons for citing a given paper" [8]. Even assuming that such a citation model exists and is well established, the task of annotating citations with their citation functions is very difficult from a cognitive point of view. First, the "citation function is hard to annotate because it in principle requires interpretation of author intentions (what could the author's intention have been in choosing a certain citation?)" [7]. Second, one has to create his/her own mental model of the citation model, so as to associate a particular meaning to each of the various functions defined by the citation model. Third, one has to map, by means of the mental model, the personal interpretation of author's intention emerging from a written text containing a citation with the one of the functions of the citation model.

Our work is positioned within the field of "Semantic Web and Cognition". In particular, the goal of this paper is to analyse weaknesses and strengths of

a particular citation model, studying how it has been used (and misused) by the users for the annotation of citations. The model under investigation is CiTO (Citation Typing Ontology)[3] [5], an OWL ontology for describing the nature of citations in scientific research articles and other scholarly works. We present the results of a preliminary user testing session with five users to whom we asked to assign CiTO properties to the citations in the Proceedings of Balisage 2011.

The paper is then structured as follows. In Section 2 we introduce previous works on classification of citations. In Section 3 we present our experimental setting and results: we go into details of the analysis performed by the humans and discuss the outcomes. Finally we conclude the paper sketching out some future works in Section 4.

2 Related works

Teufel et al. [7] [8] study the function of citations – that they define as "author's reason for citing a given paper" – and provide a categorisation of possible citation functions organised in twelve classes, in turn clustered in Negative, Neutral and Positive rhetorical functions. Jorg [3] analysed the ACL Anthology Networks[4] and found one hundred fifty cue verbs, i.e. verbs usually used to carry important information about the nature of citations: based on, outperform, focus on, extend, etc. She maps cue verbs to classes of citation functions according to the classification provided by Moravcsik et al. [4] and makes the bases to the development of a formal citation ontology.

These works actually represent some of the sources of inspiration of CiTO (the Citation Typing Ontology) developed by Peroni et al. [5], which is the ontology we used in our experiment. CiTO permits the motivations of an author when referring to another document to be captured and described by using Semantic Web technologies and languages such as RDF and OWL.

3 Using CiTO to characterise citations

In order to assess how CiTO is used to annotate scholarly articles, we compared the classifications performed by humans on a set of citations. The role of CiTO in such a process was obviously prominent. We in fact used the experiment to study the effectiveness of CiTO, to measure the understandability of its entities, and to identify some possible improvements, extensions and simplifications.

Our goal was to answer to the following four research questions (RQs):

1. How many CiTO properties have been used by users during the test?
2. What are the most used CiTO properties?
3. What is the global inter-rater agreement among users?
4. What are the CiTO properties showing an acceptable positive agreement between users?

[3] CiTO: http://purl.org/spar/cito.
[4] ACL Anthology Network: http://clair.eecs.umich.edu/aan/index.php.

The test bed includes some scientific papers encoded in XML DocBook, containing citations of different types. The papers are all written in English and chosen among those published in the proceedings of the Balisage Conference Series (devoted to XML and other kinds of markup). We automatically extracted citation sentences, through an XSLT transform, from all the papers published in the seventh volume of the proceedings, which are freely available online[5]. The XSLT transform is available at http://www.essepuntato.it/2013/citalo/xslt.

We took into account only those papers for which the XSLT transform retrieved at least one citation (i.e. 18 papers written by different authors). The total number of citations retrieved was 377, for a mean of 20.94 citations per paper. We then filtered all the citation sentences that contain verbs (extends, discusses, etc.) and/or other grammatical structures (uses method in, uses data from, etc.) that carry explicitly a particular citation function. We considered that rule as a strict guideline as also suggested by Teufel et al. [7]. We obtained 104 citations out of 377, obtaining at least one citation for each of the 18 paper used (with a mean of 5.77 citations per paper). These citations are very heterogeneous and provide us a significative sample for analysing human classifications. Finally, we manually expanded each citation sentence (i.e. the sentence containing the reference to a bibliographic entity) selecting a context window[6], that we think is useful to classify that citation.

3.1 Results

The test was carried on, through a web interface, by five users, all academic but not necessarily expert in Computer Science (the main area of the Balisage Conference). None of them was an expert user of CiTO. Each user processed each citation sentence separately, with its full context window, and had to select one CiTO property for that sentence. Users could also revise their choices and perform the experiments off-line. There was no time constraint and users could freely access the CiTO documentation. We used R[7] to load the data and elaborate the results. All the data collected are available online at http://www.essepuntato.it/2013/aic2013/test.

The experiments confirmed some of our hypotheses and highlighted some unexpected issues too. The first point to notice is that our users have selected 34 different CiTO properties over 40, with an average of 22.4 properties per user (RQ1). Moreover a few of these properties have been used many times, while most of them have been selected in a small number of cases, as shown in Table 1 (RQ2). There were 6 properties not selected by any user: compiles, disputes, parodies, plagiarizes, refutes, and repliesTo.

These data show that there is a great variability in the choices of humans. In fact only 3 citations (out of 104) have been classified with exactly the same

[5] Proceedings of Balisage 2011: http://balisage.net/Proceedings/vol7/cover.html.

[6] The context window [6] of a citation is a chain of sentences implicitly referring to the citation itself, which usually starts from the citation sentence and involves few more subsequent sentences where that citation is still implicit [1].

[7] R project for statistical computing: http://www.r-project.org/.

Table 1. The distribution of CiTO properties selected by the users.

# Citations	CITO property
110	citesForInformation
39	citesAsRelated
38	citesAsDataSource
32	citesAsAuthority, obtainsBackgroundFrom
28	citesAsEvidence, citesAsSourceDocument
24	obtainsSupportFrom
23	citesAsRecommendedReading, usesMethodIn
21	citesAsPotentialSolution
< 21	agreesWith, citesAsMetadataDocument, containsAssertionFrom, credits, critiques, discusses, documents, extends, includesQuotationFrom, usesConclusionsFrom
< 5	confirms, corrects, derides, disagreesWith, includesExcerptFrom, qualifies, retracts, reviews, ridicules, speculatesOn, supports, updates, usesDataFrom

CiTO property by all 5 users, while for 23 citations the humans selected at most two properties. These results are summarised in Table 2, together with the list of selected properties. In that table, we indicate how many citations of the dataset users agreed, and the number of properties selected by the users.

Table 2. The distribution of citations and CiTO properties on which users agreed.

Max # of properties per citation	# Citations in the dataset	CiTO properties
1	3	citesAsDataSource (5), citesAsPotentialSolution (5), citesAsRecommendedReading (5)
2	23	citesForInformation (27), citesAsDataSource (21), citesAsRelated (16), citesAsRecommendedReading (11), citesAsPotentialSolution (9), citesAsAuthority (6), credits (4), includesQuotationFrom (4), critiques (3), discusses (3), obtainsBackgroundFrom (3), usesMethodIn (3), citesAsSourceDocument (2), obtainsSupportFrom (2), citesAsEvidence (1)

3.2 Evaluation

Considering all the 104 citations, the agreement among humans was very poor. We measured the Fleiss' k (that assesses the reliability of agreement between a fixed number of raters classifying items) for the 5 raters over all 104 subjects and obtained $k = 0.16$, meaning that there exists a positive agreement between users but it is very low (RQ3). However there exists a core set of CiTO properties whose meaning is clearer for the users and on which they tend to agree. In fact, even considering the whole dataset whose k value was very low, we found a moderate positive local agreement (i.e. $0.33 <= k <= 0.66$) on some proper-

ties (RQ4): citesAsDataSource (k = 0.5), citesAsPotentialSolution (k = 0.45), citesAsRecommendedReading (k = 0.34), includesQuotationFrom (k = 0.49).

The results on the core CiTO properties were also confirmed by a slightly different analysis. We filtered only the 23 citations on which the users used at most two properties, as mentioned earlier in table Table 2. The k value on that subset of citations showed a moderate positive agreement between humans (k = 0.55, with 5 raters over 23 subjects). We had also moderate and high local positive agreement (i.e. k > 0.66) for 10 of the 15 properties used. The 5 properties showing an high positive agreement are citesAsDataSource (k = 0.77), citesAsPotentialSolution (k = 0.88), citesAsRecommendedReading (k = 0.7), credits (k = 0.74, that was not included in the core set mentioned above), and includesQuotationFrom (k = 0.74); the properties showing a moderate positive agreement are citesAsRelated (k = 0.6), citesForInformation (k = 0.4), critiques (k = 0.49), obtainsBackgroundFrom (k = 0.49), and usesMethodIn (k = 0.49).

3.3 Discussion

One of our findings was that some of the properties were used only few times or not used at all. This result can depend on a variety of factors. First, the authors of the articles in our dataset, which are researchers on markup languages, use a quite specific jargon so the citation windows resulted not easy to interpret with respect to citations. Second, the positive or negative connotation of the properties was difficult to appreciate. For instance, the fact that the properties carrying negative judgements (corrects, derides, disagreesWith, etc.) are less frequent than the others supports the findings of Teufel et al. [7] on this topic.

Although we think the intended audience of the research articles one chooses for such an experiment may bias the use of some properties, we also believe that some properties are actually shared among different scholarly domains. The property citesForInformation is a clear example. As expected, it was the most used property, being it the most neutral of CiTO. This is in line with the findings of Teufel et al. [8] on the analysis of citations within Linguistics scholarly literature, where the neutral category Neut was used for the majority of annotations by humans. Although its large adoption, citesForInformation had a very low positive local agreement (k = 0.13). This is not surprising since the property was used many times, often as neutral classification on citations that were classified in a more precise way by other users.

One of the reasons for having a low positive agreement in total (i.e. k = 0.16) could be the high number of properties (40) defined in CiTO. To test this, we mapped the 40 CiTO properties into 9 of the 12 categories identified by Teufel et al. [8][8] and re-calculated the Fleiss' k obtaining k = 0.19. Even if the agreement is slightly better than the one we got initially, the number of available choices did not impact too much. It seems to be only one of the factors to take into account for that low agreement. Another important factor might have been the

[8] The alignments of the forty CiTO properties with Teufel et al.'s classification is available at http://www.essepuntato.it/2013/07/teufel.

flat organisation of CiTO properties. Since there is no hierarchical structure, each user followed its own mental mapping and ended up selecting very different values – probably because users' mental models differed largely between users.

We also asked humans informally what were the cognitive issues they experienced during the test. Some of them highlighted that it was easy to get lost in choosing the right property for a citation because of the large number of possible choices. In addition, they also claimed that supporting the documentation of CiTO with at least one canonical example of citation for each property could be useful to simplify the choice.

4 Conclusions

The main conclusion for this paper is that classifying citations is an extremely difficult job also for humans, as demonstrated in our experiments on the properties of CiTO. The human analysis we presented herein gave us important hints on the understanding and adoption of CiTO, still showing some uncertainty and great variability. The identified strengths and weaknesses will be used to further improve the ontology, together with experiments on a larger set of users, decreasing the number of possible choices (for instance by using only the CiTO properties showing more agreement among humans).

References

1. Athar, A., Teufel, S. (2012). Detection of implicit citations for sentiment detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: 18-26.
2. Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2013). Semantic Annotation of Scholarly Documents and Citations. To appear in Proceedings of 13th Conference of the Italian Association for Artificial Intelligence (AI*IA 2013).
3. Jorg, B. (2008). Towards the Nature of Citations. In Poster Proceedings of the 5th International Conference on Formal Ontology in Information Systems.
4. Moravcsik, M. J., Murugesan, P. (1975). Some Results on the Function and Quality of Citations. In Social Studies of Science, 5 (1): 86-92.
5. Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 17 (December 2012): 33-43. DOI: 10.1016/j.websem.2012.08.001
6. Qazvinian, V., Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: 555-564.
7. Teufel, S., Siddharthan, A., Tidhar, D. (2006). Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing: 103-110.
8. Teufel, S., Siddharthan, A., Tidhar, D. (2009). An annotation scheme for citation function. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue: 80-87.