Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2009 November 1; 1-4 Nov(2009): 308–313. doi: 10.1109/BIBMW.2009.5332080.

Investigating and Annotating the Role of Citation in Biomedical Full-Text Articles

Hong Yu, PhD, Shashank Agarwal, MS, and Nadya Frid, MS
University of Wisconsin-Milwaukee

Abstract

Citations are ubiquitous in scientific articles and play important roles for representing the semantic content of a full-text biomedical article. In this work, we manually examined full-text biomedical articles to analyze the semantic content of citations in full-text biomedical articles. After developing a citation relation schema and annotation guideline, our pilot annotation results show an overall agreement of 0.71, and here we report on the research challenges and the lessons we've learned while trying to overcome them. Our work is a first step toward automatic citation classification in full-text biomedical articles, which may contribute to many text mining tasks, including information retrieval, extraction, summarization, and question answering.

1. Introduction

A citation is defined as a *cited* article that is referenced by a *citing* article. For instance, Example 1 shows a sentence from a citing article which cites the articles (1, 2).

Example 1: "Ependymal cells form a continuous, heterogeneous monolayered epithelium that lines the ventricular surface of the vertebrate brain and central canal of the spinal chord, separating the cerebrospinal fluid (CSF) from the brain parenchyma (1,2)." From [1].

Citations are ubiquitous in scientific articles. For instance, in the TREC Genomics Track text collection [2], which incorporates more than 160,000 full-text biomedical articles, we found an average of 34 citations in each article.

In biomedical literature, a citation may introduce background knowledge and related work, or interpret results and provide supports to infer new research hypothesis. A citation plays an important role for representing the semantic content of a full-text biomedical article, and as a result, citation information has been used to assist in biomedical text mining tasks (e.g., [3-8]). For example, two articles can be considered as "related" if they share a significant set of co-citations, and a recent study incorporating this model has shown it can improve information retrieval [3]. The number of times a citation is cited in a paper may indicate its relevance to the citing paper [4,5]. Citances (that is, citation sentences) represent the condensed semantic content of the documents they identify [6,7] and can be used to extract scientific fact [8] and for summarization tasks [7].

Citation indexing has been used to associate citing articles with cited ones, and the associations have been used to create a Science Citation Index to measure the impact factor of scientific journals and articles [9]. Impact factor is used as a proxy for the importance of a journal or an article to its field [10]. However, citation indexes may be refined by providing the information on how an article is cited, whether the citing author argues or agrees with the cited one or whether knowledge or methods are cited.

A citation relation is a relation between two cited articles, or a relation between the citing article and the work cited. Example 1 shows that the cited articles “1, 2” provide the background information to the citing article. Example 2 shows a relation (in this case, a conflict relation) between two cited articles.

Example 2: “When the first mutants defective in subunit export were identified, it was assumed that the corresponding proteins functioned as components of the ribosomal subunit export machinery (Stage-Zimmermann et al. 2000). However, as the numbers of such “export factors” has grown, it has become increasingly unlikely that they all function directly in subunit export (Stage-Zimmermann et al. 2000; ...” From [11].

There are many different reasons why an author decides to include a citation (e.g., the work of [6,12-24]), and recognizing citation relations may benefit many text mining tasks. For instance, the positive evaluation presented by the authors in Example 3 (i.e., “groundbreaking work”) could be used to refine citation indexes. Example 4 shows that a cited article can be used to provide the protocol of an experiment, which is in contrast to a case in which the cited article or articles are used for background knowledge, as in Example 1. Distinguishing between these different reasons for citing may assist in the tasks of information extraction and question answering. Example 5 shows a conflict relation between the citing article and the article cited, which, if captured, may benefit the task of fact extraction.

Example 3: “The groundbreaking work of Myers et al. (34) shows that experimental m values are proportional to the change in surface area on urea-induced protein unfolding.” From [25].

Example 4: “Ependymal primary cultures were prepared according to the protocol of Prothmann et al. (17), which uses a simplified version of a previously described method (60).” From [1].

Example 5: “The ability of RAD52-(218–418) to self-associate was unexpected. Previous studies have suggested that residues 65–165 define the exclusive self-association domain in the RAD52 protein (18).” From [26].

Our long term goal is to automatically identify citation relations and to incorporate the relations to improve biomedical text mining. This work reports our first step toward this goal. Toward this end, we investigated and classified citation relations in full-text biomedical articles and developed a citation relation schema and an annotation guideline to assign each citation to its corresponding relations. We then conducted a pilot annotation on seven full-text biomedical articles. Here we report on the development of the annotation guideline, the annotation agreement, and the challenges that presented themselves in this development.

2. Background

The literature that categorizes the relations between the cited and the citing articles is rich [6,12-24]. Garfield (1965) pioneered 15 reasons why an author introduces a citation [6], including *paying homage to pioneers*, *paying homage to peers*, *providing background reading*, and *criticizing previous work*. Moravcsik and Murugesan (1975) [20] classified the role of a citation along multiple dimensions, including *conceptual* (is the citation made in connection with a concept or theory?) or *operational* (is it made in connection with a technique used in the citing article?), and *organic* (is the citation truly needed for the understanding of the citing article?) or *perfunctory* (is the citation an acknowledgment that some other work in the same general area has been performed?). Chubin and Moitra (1975) [27] classified cited papers into several categories, including *basic essential citation* (the

cited paper is declared central to the reported research; the reported findings depend on the cited paper), *subsidiary essential citation* (the cited paper is not directly connected to the subject of the letter or article but is still essential to the reported research), *partial negational citation* (the cited paper is erroneous in part, and the author of the citing paper offers a correction), and *total negational citation* (the cited paper is completely wrong, and the author of the citing paper offers an independent interpretation of the solution). Spiegel-Rosing (1977) [21] defined multiple general categories, including that the “cited source is positively evaluated.” Hanney et al. (2005) [23] divided citations into four categories: *limited*, *peripheral*, *considerable*, and *essential*. See the review article [24] for details.

On the other hand, the work in automatic citation classification is limited. Garzone and Mercer [28-30] treated the citation classification as a task of sentence categorization. They extracted a sentence that incorporated citations and then applied manually curated lexical and grammar rules to assign the citations to one of 35 predefined categories. These categories included “citing work totally disputes some aspect of cited work,” “citing work totally confirms cited work,” and “citing work refers to assumed knowledge which is general background.” More recently, Radoulov [31] built upon the work of Garzone and Mercer [28-30] to develop supervised machine-learning approaches for automatic citation classification. The citation schema they built including *reason for citation* and *what is cited*. *Reason for citation* further incorporates such sub-categories as *confirms*, *supports*, *interprets results*, *extends model*, *contrasts*, *future research*, and *uses*. *What is cited* includes the sub-categories *general background*, *specific background*, *historical account*, *pioneering work*, *related work*, *concept*, *method*, *product*, and *data*. They report a performance of ~70% f-score for automatic citation classification, although some features (e.g., cue words) were manually identified.

It is unclear whether the two fine-grained citation schemas described above lead to consistent annotations. Some of the categories defined in [31] seem to overlap. For example, *Confirms* seems to overlap with *Supports*. *General Background* may be hard to distinguish from *Specific Background*. Furthermore, the generalizability of both automatic citation classifiers described above has not yet been determined.

3. Methods

Our citation relation identification is conducted with the goal of improving biomedical information retrieval. Most existing work defines a citation relation as a relation between a cited article and its citing article. In this study, we defined a citation as a relation not only between a cited article and its citing article, but also between cited articles; example 2 shows a *conflict* relation between cited articles.

To identify and define specific citation relations, we studied the citation categories proposed in previous studies [6,12-24,28-31] before examining and annotating seven randomly selected full-text biomedical articles. We then refine this top-down citation schema through a bottom-up approach. Specifically, through an iterative process, we identified linguistic features, refined citation relations, and developed an annotation guideline for annotating citation relations. At each iteration, a biologist annotator (Agarwal) and a linguist annotator (Frid) independently annotated one full-text biomedical article. We measured overall agreement, which is defined as the percentage of agreed relations out of the total relations identified by the two annotators. We counted a group of co-occurring citations as one, rather than separating it into citation instances. For instance, “1,2” in Example 1 represents one citation.

4. Citation Relation Schema

Figure 1 shows the citation relation schema. We define three different relations: 1) *between the citing and the cited articles* (e.g., Example 1), 2) *between two cited articles* (e.g., Example 2), or 3) *between two propositions in one cited article*. Example 6 illustrates a case of *relation between two propositions in one cited article* as marked by the cue phrase “however.”

Example 6: “In that study he stated that those five species (*E. bennetti*, *E. helvola*, *E. luteola*, *E. seabrai*, and *E. tenuifasciata*) were very similar to each other and could be distinguished only by subtle differences in coloration (if any) and geographic distributions. However, in the same paper, he found that *E. helvola* and *E. luteola* share a synapomorphy (a shallow notch on the arms of the seventh sternum) which made them appear as sister groups, while *E. seabrai*, *E. bennetti* and *E. tenuifasciata* possess another apomorphic state of this character (a deep notch) and group together as another branch inside the leucopyga species-subgroup.” From [32].

A citation can represent *knowledge* (e.g., Example 1) or *an experimental protocol* (e.g., Example 4). A citation may incorporate author's comments. With respect to author's comments, we further defined categories including *reported speech and discovery*, *epistemic modality*, *evaluation* (the subcategories are *positive*, *quantitative*, and *first to discover*), *adversative* (the subcategories are *contrast*, *conflict*, *specify*, and *obstacle*), *consistency* and *cause*.

Reported speech

Is identified by attribution verbs (e.g., “report” and “describe”) and verbs of discovery (e.g., “identify” and “discover”), an example is shown in 7.

Example 7: “In 1990, Fearon and Vogelstein (1) reported that at least nine chromosomal alterations were found in up to 50% of colon adenocarcinomas in humans.” From [33].

Epistemic modality

Indicates the degree of the speaker's certainty that the proposition is true. Examples of modality markers are modal verbs (e.g., “can,” “may,” “might”), attribution verbs (e.g., “hypothesize”) and adverb (e.g., “probably”).

Evaluation

Is assigned if the author of the citing paper makes an evaluative remark (*positive* as shown in Example 3). Lexical cues such as “numerous studies” or “extensive research” identify *quantitative evaluation*, as shown in Example 8. An example of *first to discover* is 9.

Example 8: “Numerous studies have explored possible routes of delivering cells for engraftment in the hepatic parenchyma (17)” From [34].

Example 9: “Schwab and his colleagues (1) were the first to discover a potent growth inhibitory activity in the CNS, and to show that much of this inhibitory activity is associated with myelin.” From [35].

Adversative

Include *contrast*, *conflict*, *specify*, and *obstacle*. *Contrast* indicates that one has a property, while the other does not have. *Contrast* is assigned to contrasting features of different objects. *Conflict* assigns opposing opinions. *Specify* (e.g., Example 10) means that the citing

author (or another cited author) does not entirely deny but rather specifies what is claimed in the cited paper, i.e., states under which conditions the cited idea is true. *Obstacle* (e.g., Example 11) indicates that a situation in a cited article faced some obstacle.

Example 10: “When the first mutants defective in subunit export were identified, it was assumed that the corresponding proteins functioned as components of the ribosomal subunit export machinery (Stage-Zimmermann et al. 2000). However, as the numbers of such “export factors” has grown, it has become increasingly unlikely that they all function directly in subunit export....” From [11].

Example 11: “We and others argued that it should be possible, in principle, to use digital PCR to create a universal, polymorphismindependent test for fetal aneuploidy by using maternal plasma DNA (17–19), but because of technical challenges relating to the low fraction of fetal DNA, such a test has not yet been practically realized.” From [36].

Consistency

Indicates similarity in results or protocol. *Cause* value is used when a proposition in the cited text serves as a cause for the proposition in the citing text. Examples of two categories are shown below, respectively.

Example 12: “These findings suggest that the majority of cell-free DNA in the plasma is derived from apoptotic cells, in accordance with previous findings (22, 23, 25, 26).” From [36].

Example 13: “Because islets are particularly prone to injury during inflammatory conditions (4), the functioning islet mass rapidly decreases after transplantation before antigenic recognition (5).” From [37].

5. Annotation Results

We annotated a total of seven full-text biomedical articles. The number of citation relations that were assigned *between cited* or *between citing and cited* in seven articles was 404, and the two annotators assigned 1,428 relations in total. Figure 2 shows the distribution of citation relations in those seven articles. We did not find cases of certain citation relations (e.g., *relation between two propositions in one cited article*). Out of the total 1,428 assigned relations, two annotators agreed on 1,013. The overall pairwise agreement was 0.71. Figures 3 and 4 show the annotation agreement in seven articles that were annotated over time. The results show that the best agreement (>0.9) was attained when identifying whether a relation was between the citing and the cited articles or between two cited articles. As shown in Figure 4, the agreement of annotators on *relations between the citing and the cited* has the highest agreement (close to 1), while the agreement of *relations between the cited* can be blurred (the agreement was 0 in 4 out of 7 articles).

We found a consistent annotation agreement in relation to knowledge (>0.75). The agreement in relation to experimental protocol ranges from 0.4 to 0.8.

The category of *author's comments* yields the worst agreement (0.4-0.7). When we broke down each subcategory, we found that *no author's comments*, *reported speech and discovery* and *consistency* yielded more consistent and better agreement (0.2-0.9) than *cause*, *modality*, *positive evaluation*, *contrast*, *quantitative evaluation*, *specify* and *conflict*.

6. Discussion and Future Work

As shown in Figure 4, we found that the overall annotation agreement of *between the citing and cited* was consistently high (>0.9), while the agreement of *between cited* was low. The reasons were threefold: (1) the number of relations was low (26 out of the total 404 citation relations), (2) the annotation schema may need to be further refined, and (3) the annotation requires the understanding of discourse relations, which is a challenging task because of the ambiguity of biomedical text.

In comparison with Garzone and Mercer [28-30] or Radoulov's citation schema [31], our citation schema is more simplified and intuitive, and therefore easier to annotate. Even so, we found that annotating certain relations (e.g., *cause* and *quantitative evaluation*) suffered from a low number of occurrences and poor agreement, even after successive iterations. We speculate that agreement could be improved as we annotate more articles and continue to refine the annotation schema.

Figure 3 shows that the agreement of annotators on parameter A, which establishes relation between citing and cited articles, was high (close to 1), while the agreement for parameter C, which concerns author's attributions, was low (about 0.5). The results also show that the annotations in most *author's comments* subcategories were challenging. For example, Example 13 was annotated by NF as *specify* and SA as *contrast*. Again, we plan to further simplify the annotation schema to improve consistency. In addition, we plan to rely on linguistic cues for determining relations.

Example 13: "Although AAT inhibits caspase-3 in a rat beta cell line (39), an antiapoptotic function of AAT per se is probably incapable of inducing tolerance, because global inhibition of apoptosis during an inflammatory and/or immune response would promote an expanded reactive T cell population. We therefore suggest that AAT can be considered for testing in human islet transplantation." From [37].

Citation scope (i.e., the exact boundaries of a citation) is also a challenging task in annotation. In our annotation, we assign the citation sentence as the scope of the corresponding citation. However, such simplification has its cost. In many cases, the citation scope is ambiguous. Example 14 shows that the scope can be the whole sentence or just a proposition (brackets stand for citation boundaries). In future work we will continue to refine our annotation guideline, conduct evaluation on a larger scale, and resolve such challenges for annotation, as ambiguity and scope. It is our long term goal to develop an automatic citation relation identification system.

Example 14: "[Bmi-1, a polycomb transcription factor, regulates neural stem cell self-renewal *in vivo* at postnatal, but not prenatal, stages, probably because [fetal neural stem cells do not express the cell cycle inhibitory proteins that are regulated by Bmi-1 *in vivo*] (35)." From [38].

Acknowledgments

The authors acknowledge support from the National Library of Medicine to Hong Yu, grant number 1R01LM009836-01A1. We thank Lamont Antieau for editing the paper.

References

1. Wellard J, Rapp M, Hamprecht B, Verleysdonk S. Atrial natriuretic peptides elevate cyclic GMP levels in primary cultures of rat ependymal cells. *Neurochem Res* 2003;28:225-233. [PubMed: 12608696]

2. Hersh, W.; Cohen, A.; Roberts, P.; Rekapalli, H. TREC 2006 Genomics Track overview. TREC Genomics Track conference; 2006.
3. Tbahriti I, Chichester C, Lisacek F, Ruch P. Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library. *Int J Med Inform* 2006;75:488–495. [PubMed: 16165395]
4. Voos H, Dagaev K. Are all citations equal? Or, did we op. Cit. your Idem? *Journal of Academic Librarianship* 1976;1:19–21.
5. Herlach G. Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science* 1978;29:308–310.
6. Garfield, E. Can citation indexing be automated?. In: Stevens, Mary Elizabeth; G, VE.; Heilprin, Laurence B., editors. *Statistical Association Methods for Mechanized Documentation*. Washington. Symposium Proceedings: Volume 269 of National Bureau of Standards Miscellaneous Publication. 1965. p. 189-192.
7. Schwartz, A.; Hearst, M. Summarizing key concepts using citation sentences. *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language Processing and Biology*; 2006. p. 134-135.
8. Nakov, P.; Schwartz, A.; Hearst, M. Citances: Citation sentences for semantic analysis of bioscience text. *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*; Sheffield, UK. 2004.
9. Garfield, E. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. New York: Wiley; 1979.
10. Garfield E. Citation analysis as a tool in journal evaluation: journals can be ranked by frequency and impact of the citations for science policy studies. *Science* 1972;178:471–479. [PubMed: 5079701]
11. Thomson E, Tollervey D. Nop53p is required for late 60S ribosome subunit maturation and nuclear export in yeast. *Rna* 2005;11:1215–1224. [PubMed: 16043506]
12. Frost C. The use of citations in literary research: A preliminary classification of citation functions. *Library Quarterly* 1979;49:399–414.
13. Lipetz B. Problems of citation analysis: critical review. *American Documentation* 1965;16:381–390.
14. Cole, S. *The growth of scientific knowledge: theories of deviance as a case study*. New York: Harcourt Brace Jovanovich; 1975.
15. Peritz B. A classification of citation roles for the social sciences and related fields. *Scientometrics* 1983;5:303–312.
16. Small H. Cited documents as concept symbols. *Social Studies of Science* 1978;8:327–340.
17. Weinstock, M. *Citation indexes*. New York: Marcel Dekkar; 1971.
18. Finney, B. MS thesis. The City University of London; 1979. The reference characteristics of scientific texts.
19. Duncan, E.; Anderson, F.; McAleese, R. *Information Retrieval in Educational Technology: Proceedings of the First Symposium on Information Retrieval in Educational Technology*. University of Aberdeen; 1981. Qualified citation indexing: its relevance to educational technology; p. 70-79.
20. Moravcsik M, Murugesan P. Some results on the function and quality of citations. *Social Studies of Science* 1975;5:86–92.
21. Spiegel-Rosing I. Science studies - bibliometric and content-analysis. *Social Studies of Science* 1977;7:97–113.
22. Krampen, G.; Montada, L. *Wissenschaftsforschung in der Psychologie*. Hogrefe; Gottingen, Germany: 2002.
23. Hanney, S.; Grant, J.; Jones, T.; Buxton, M. *Categorising citations to trace research impact In Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*. Stockholm, Sweden: Karolinska University Press; 2005.

24. Bornmann L, Daniel H. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*. 2007 in press.
25. Auton M, Holthauzen LM, Bolen DW. Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc Natl Acad Sci U S A* 2007;104:15317–15322. [PubMed: 17878304]
26. Ranatunga W, Jackson D, Lloyd JA, Forget AL, Knight KL, Borgstahl GE. Human RAD52 exhibits two modes of self-association. *J Biol Chem* 2001;276:15876–15880. [PubMed: 11278978]
27. Chubin D, Moitra S. Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science* 1975;5:423–441.
28. Garzone, M.; Mercer, R. Towards an automated citation classifier. Proceedings on 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence; Montreal, Quebec, Canada. 2000.
29. DiMarco, C.; Mercer, R. Pacific Association for Computational Linguistics (PACLING 2003). Halifax, Canada: 2003. Toward a catalogue of citation-related rhetorical cues in scientific texts.
30. DiMarco, C.; Mercer, R.; Rubin, V. A design methodology for a document indexing tool using pragmatic evidence in text. Proceedings of the 3rd Annual Scientific Conference of Intelligent, Interactive, Learning Object Repositories Network (LORNET Research Network); Montreal, Canada. 2006.
31. Radoulov, R. Computer Science. University of Waterloo; 2008. Exploring automatic citation classification; p. 102
32. Nemesio A, Silveira FA. First record of *Eulaema helvola* Moure (Hymenoptera: Apidae: Euglossina) for the State of Minas Gerais: biogeographic and taxonomic implications. *Neotrop Entomol* 2006;35:418–420. [PubMed: 18575707]
33. de Boer JG. A new mutator phenotype in breast cancer? *Proc Natl Acad Sci U S A* 2002;99:3368–3369. [PubMed: 11904400]
34. Fair JH, Cairns BA, Lapaglia MA, Caballero M, Pleasant WA, Hatada S, Kim HS, Gui T, Pevny L, Meyer AA, Stafford DW, Smithies O, Frelinger JA. Correction of factor IX deficiency in mice by embryonic stem cells differentiated in vitro. *Proc Natl Acad Sci U S A* 2005;102:2958–2963. [PubMed: 15699326]
35. McKerracher L. Ganglioside rafts as MAG receptors that mediate blockade of axon growth. *Proc Natl Acad Sci U S A* 2002;99:7811–7813. [PubMed: 12060723]
36. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci U S A* 2008;105:16266–16271. [PubMed: 18838674]
37. Lewis EC, Mizrahi M, Toledano M, Defelice N, Wright JL, Churg A, Shapiro L, Dinarello CA. alpha1-Antitrypsin monotherapy induces immune tolerance during islet allograft transplantation in mice. *Proc Natl Acad Sci U S A* 2008;105:16236–16241. [PubMed: 18852465]
38. Groszer M, Erickson R, Scripture-Adams DD, Dougherty JD, Le Belle J, Zack JA, Geschwind DH, Liu X, Kornblum HI, Wu H. PTEN negatively regulates neural stem cell self-renewal by modulating G0-G1 cell cycle entry. *Proc Natl Acad Sci U S A* 2006;103:111–116. [PubMed: 16373498]

- A.
 - A1. Relation between the citing and the cited articles
 - A2. Relation between two cited articles
 - A3. Relation between two propositions in one cited article
- B.
 - B1. Related to knowledge
 - B2. Related to experimental protocol
- C.
 - C1 No author's comments
 - C2 Author's comments:
 - C2a reported speech and discovery
 - C2b epistemic modality
 - C2c evaluation
 - C2c1 positive evaluation
 - C2c2 quantitative evaluation
 - C2c3 first to discover
 - C2d adversative
 - C2d1 contrast
 - C2d2 conflict
 - C2d3 specify
 - C2d4 obstacle
 - C2e consistency
 - C2e1 similar features
 - C2e2 similar opinions, results and methods
 - C2f cause

Figure 1. Citation Relation Schema

Figure 2.
The distribution of annotated citation relations in seven articles.

Figure 3.
The overall agreement in A, B, C categories specified in Figure 1.

Figure 4.
The overall agreement in different subcategories specified in Figure 1.