Automated Citation Sentiment Analysis: What Can We Learn From Biomedical Researchers

Bei Yu
Syracuse University
Syracuse, New York, 13244
byu@syr.edu

ABSTRACT

Automated citation sentiment analysis is a newly emerged research topic inspired by traditional citation context analysis in scientometrics and applied linguistics. The main goals of current citation sentiment analysis are to develop new tools to model scientific literature and provide authoring support for researchers in tasks like literature review. In terms of developing authoring support tools, however, current studies have not taken into consideration the behavioral patterns of researchers' literature review practice, leaving user need assessment as a missing piece in current interdisciplinary research effort. This paper analyzed biomedical researchers' need by reviewing their publications on using manual citation sentiment analysis for detecting citation bias, and discussed the differences between biomedical researchers' approach and current automated citation sentiment analysis model. These differences are expected to inform the modeling of automated citation sentiment analysis.

Keywords

Citation, sentiment analysis, bibliometrics, scientometrics, citation bias, user study, need assessment.

INTRODUCTION

Automated citation sentiment analysis is a newly emerged research topic in the field of natural language processing (e.g. Teufel et al., 2006; Schafer & Spurk, 2010; Small, 2011; Athar & Teufel, 2012). An automated system is expected to use machine learning algorithms and linguistic cues to identify a citing author's opinion toward the cited work as expressed in the citation context. Once the sentiments of all citations to a cited work were identified, a comprehensive evaluation of the cited work would emerge. This comprehensive evaluation may help scientometrics scholars and research administrators assess the cited work's research contribution and impact, which is currently conducted based on citation counts. Automatic citation sentiment identification may also provide scientific authoring support by helping researchers with thorough literature search and review when the volume of scientific literature keeps increasing rapidly. This paper focuses on using citation sentiment analysis for scientific authoring support (Nanba et al., 2004; Qazvinian & Radev, 2008; Zhang et al., 2008; Ritchie et al., 2008; Shafer & Kasterka, 2010).

Among the pioneering studies on automated citation sentiment analysis, some defined citation sentiment as the polarity of a citing author's opinion (e.g. Athar & Teufel, 2012). The polarity is usually categorized as positive, negative or neutral, in accordance with the typical definition of a sentiment classification task (Pang & Lee, 2008). Other studies defined citation sentiment as a fine-grained classification of citation functions, which can consist of many more categories than the polarity based definition (Shafer & Spurk, 2010; Small, 2011). The fine-grained citation sentiment definition is similar to the traditional citation function analysis in scientometrics (e.g. Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975). For differentiating purpose, this paper adopts the polarity based citation sentiment definition, and refers to the fine-grained definition as citation function analysis.

In order to design automated citation sentiment analysis tools for scientific authoring support, researchers as the end users should be involved in the system design process; their needs should be assessed to inform user-centered design. However, prior studies were mainly designed based on either authors' own research experience, which can be subjective, or traditional scientometrics studies, which do not aim for scientific authoring support. In consequence, these automated citation sentiment analysis studies have not taken into consideration the behavioral patterns of researchers' literature review practice, leaving user need assessment as a missing piece in current interdisciplinary research effort.

This paper takes a non-obtrusive retrospect approach to analyzing biomedical researchers' needs for automated citation sentiment analysis tool by reviewing their publications on using manual citation sentiment analysis for detecting citation bias. Citation bias refers to the phenomenon that negative results, including insignificant findings and dissenting opinions, received much fewer citations than positive results, leading the research community to a distorted view of existing literature. Citation bias has been reported in many different biomedical studies (e.g. Greenberg, 2009; Fiorentino et al., 2011; Shrag et al., 2011). To detect citation bias toward a specific scientific claim, biomedical researchers have been conducting their own citation sentiment analyses by systematically searching relevant papers, manually reading the citation contexts, identifying whether they support or oppose the specific claim, and finally plugging their manual annotations into the citation network to visualize the citation bias. This claim-specific citation sentiment approach differs from the current automated methods that focus on the relationship between citers and citees rather than citers and claims. Driven by their own need for detecting citation bias, biomedical researchers have designed a different citation sentiment analysis approach that overlaps with the current automated citation sentiment analysis approach. This paper aims for identifying the differences between the two approaches, which are expected to inform better user-centered design for automated citation sentiment analysis.

This paper is organized as follows: we will first review current studies on automated citation sentiment analysis and relevant scientometrics and linguistics studies that inspired the automated analysis. We will then move on to review biomedical researchers' manual citation sentiment analysis, and compare the methodological differences between biomedical researchers' approaches and the current automated analyses. Finally, we will discuss how these differences can inform the automated analyses as useful tools for scientific authoring support.

AUTOMATED CITATION SENTIMENT ANALYSES

An illustrative example

We use citations to (Hayes et al., 1990) as an example to illustrate the automated citation sentiment analyses process. This article presented *CONSTRUE*, an automated text categorization system. This article had accumulated 107 citations in Google Scholar as of April 7th, 2013.

The first step is to establish the ground truth of citation sentiment by manually annotating a corpus. The unit of analysis is a citation statement, defined as a block of context that involves a particular citation. A citation statement can be as short as a sentence, or span across multiple sentences or even paragraphs. Citation sentiment is annotated for each statement. Table 1 sampled five citation statements that hold conflicting opinions. By common understanding of polarity, #1 is clearly negative, questioning the test data's representativeness. This citation statement not only spans three sentences, but also contains another nested statement - the positive citation toward (Yang, 1999). #2 also criticized the data representativeness, but the negative comment was mitigated by starting with praise. #3 seems neutral since no linguistic cues indicated positivity or negativity; however, it is also reasonable to infer that #3 is implicitly positive, since it trusted the cited work by using it as a benchmark system. #4 is clearly positive, praising *CONSTRUE* as one of the successful text categorization systems. #5 also seems neutral without explicit cues of polarity. However, it may also be considered as undefined as in (Shafer & Spurk, 2010) because the citation statement did not explicitly explain the relationship between the citing and cited papers, making the judgment difficult.

Citation opinion	Citation statement
Negative	"**Hayes et al. [1990]** reported a .90 "breakeven" result (see Section 7) on a subset of the Reuters test collection, a figure that outperforms even the best classifiers built in the late '90s by state-of-the-art ML techniques. However, no other classifier has been tested on the same dataset as Construe, and **it is not clear** whether this was a randomly chosen or a favourable subset of the entire Reuters collection. **As argued in [Yang 1999]**, the results above **do not allow us to state** that these effectiveness results may be obtained in general.
Negative (mitigated)	"A **well-known** example of an expert system for this task is the CONSTRUE system **[Hayes et al. 1990]** used by the Reuters news service. … While these **are exceptionally good** results, **the test set seems to have been relatively sparse** when compared to the number of possible topics. "
Neutral (implicitly positive)	"As comparison, we use an existing text categorization system, TCS, developed using a text categorization shell built by Carnegie Group **[Hayes et al., 1990]."**
Positive	"Various **successful** systems have been developed to classify text documents including telegraphic messages [Young][Goodman], physical abstract [Biebricher], and full text news stories **[Hayes]**[Rau]."
Neutral	"The training documents can be used by human experts to generate categorization rules ([1], **[7]**) or …"

Table 1. Citation sentiment toward (Hayes et al., 1990)

The above example demonstrates the importance of automated citation sentiment analysis in assisting researchers' routine literature review tasks. In this case, the cited work received mixed feedback, and a full examination is needed to prevent undesirable bias toward prior studies. Assuming it takes 5 minutes to download and examine one citing paper, a full examination of the 107 citations would take at least nine hours. In reality, researchers can only afford to examine a small portion of the citations, usually the top ones ranked by the bibliographic systems by criteria like citing papers' citation count, publication year, or content relevance.

This example also shows that the ground truth of citation sentiment may subject to annotator's subjectivity. Teufel et al. (2006a) used three annotators to independently code 26 computational linguistics papers with 548 citation statements, and reported the inter-coder agreement at .75 Kappa value, indicating adequate level of agreement so as to use this corpus for training and evaluating sentiment classification algorithms in the next step. Teufel et al. also found neutral citations account for the majority, and polarized citations, especially negative citations, are rare.

Various classical text classification algorithms have been used for citation sentiment classification, such as nearest-neighbor algorithm in (Teufel et al., 2006b), Support Vector Machines in (Athar, 2011), and rule-based method in (Shafer & Spurk, 2010). Different feature sets have been tested on computational linguistics papers. Athar (2011) reported .764 macro-F value using SVM and a feature set consisting of bag of words, negations, and dependency relations (Marneffe & Manning, 2008). Athar & Teufel (2012) identified more polarized citations, especially negative citations, after expanding the range of citation context to sentences before and after the citing sentence, indicating the importance of identifying correct boundaries of citation statements (O'Connor, 1982; Schwartz et al., 2007; Angrosh et al., 2010). These experiments were mainly conducted on computational linguistics papers. Given the significant disciplinary difference in scientific opinion expression (Hyland, 1999), their generalizability to other discipline remains an open question. Since this paper focuses on modeling the task of citation sentiment classification, extensive review on technical details is not pursued.

Root in scientometrics and academic writing studies

Citation behavior has been extensively studied in the field of scientometrics in order to develop appropriate evaluation tools to assess research contributions. Most studies used citation count as a quantitative measure (Garfield, 1979; Bornmann and Daniel, 2008). Considering the limited information that citation count carries (MacRoberts & MacRoberts, 1989), some studies further examined citation context (McCain & Turner, 1989; Bormann & Daniel, 2008) and citers' motives (Cronin, 1984). Extensive review of this field is beyond the scope of this paper. Here we focus on prior studies that are closely relevant to citation sentiment analysis.

One task in citation context analysis is to create a typology of citation functions, and a number of complicated citation function classification schemes had been proposed since the 1960s. See (Peritz, 1983) and (Bonzi, 1982) for reviews of those classification schemes. These schemes were constructed based on publications from different disciplines, and lacked consensus on the names and definitions of the categories (Baldi, 1998). This caused trouble for establishing ground truth to train and evaluate automated methods. Actually, NLP researchers had tried to automate citation function classification based on these prior schemes, but had to revise or consolidate them due to their complexity and discrepancies, resulting in more schemes (Garzone & Mercer, 2000; Teufel et al., 2006a). No user studies have been conducted to examine whether these fine-grained classification is actually needed for assisting researchers in literature review.

However, these idiosyncratic classification schemes did share at least two things in common. First, a large proportion of citations are considered *"perfunctory"*, that is, the cited work does not substantially contribute to the citing work, compared to the rest *"organic"* citations. (e.g. Chubin and Moitra, 1975; Moravcsik and Murugesan, 1975). Inter-coder agreements were not reported in these studies. Later, Agarwal et al. (2010) reported .49 Kappa value, indicating subjectivity in the definitions.

Second, citation sentiment is a common dimension in these classification schemes. It was defined as "*questioned*", "*affirmed*", and "*refuted*" in Lipetz's 29-category system, which also included other categories like *"reviewed or compared"*, *"applied"*, and *"improved"* (Lipetz, 1965). Citation sentiment categories were also named as "*confirmative*", "*negational*", and "*neither*" in (Moravcsik and Mururgesan, 1975), "*afffirmative*" and "*negational*" in (Chubin and Moitra, 1975), and "*corroborative*", "*oppositional*", and "*corrective*" in (Hodges, 1972). These categories were basically the same as our sentiment polarity categories.

Applied linguists were also interested in analyzing the linguistic characteristics of scientific criticism (Swales, 1986), such as using reporting verbs to construct scientific arguments (Thompson & Ye, 1991), using hedges and mitigated negations to express critical comments (MacRoberts and MacRoberts, 1984; Hyland, 1994), and different distributions of negative citations in different disciplines with fewer negative citations in hard science publications (Hyland, 1999). The studies in scientometrics and academic writing are mutually informative, which has been well documented in a citation analysis in (White, 2004). Their findings have helped selecting relevant features or rules for automated citation sentiment classification.

However, citation sentiment was never addressed as a tool to assist researchers' literature review in these studies, which aimed for developing assessment methods to appropriately evaluate researchers and their publications' contributions for research administration purpose, or to teaching scientific writing to students, especially non-native English speakers. For scientometrics, the envisagement of using an arbitrary automated tool to identify citation sentiment seems too dangerous if the result was going to affect promotion and grant decisions, while the manual analysis cost is too high. This conjecture is consistent with the concerns of difficulty expressed in scientometrics studies. For example, Peritz (1983) wrote that "the nuances and gradations between affirmation and negation are so varied as to defy classification". White (2004) concurred that this task may not be delegated to computer because it requires close reading, domain knowledge, and expert judgment to apply. Therefore, if an automated citation sentiment analysis tool is designed for research administrators, they would have to require very high accuracy to be able to use it.

However, if the tool is designed as an enhanced function for current literature search tools, researchers would have higher tolerance to the classification errors, given the list of indistinguishable citations they can get from current bibliographic databases. If the tool is designed with interactive functions, researchers might even be willing to invest their precious time to contribute their feedback to improve the automated systems, just as email users are willing to correct the mistakes that spam filters make (Gormack & Lynam, 2007). Since the majority of citations are neutral, the automated tool can also be designed to allow users to adjust the balance between precision and recall – a tight setting may find fewer polarized citations but also raise fewer false alarms, and a loose setting may improve the coverage with the cost of returning more false alarms.

BIOMEDICAL RESEARCHERS' MANUAL CITATION SENTIMENT ANALYSES

A non-obtrusive, retrospect approach for user study

In order to design automated citation sentiment analysis tools for scientific authoring support, we need to investigate whether and how researchers as the end users can be helped by citation sentiment analysis. Unfortunately, user need assessment has been missing in current studies. One possible reason is that the system designers are researchers themselves, who may designed the system based on their own experience. However, the numerous disparate citation classification schemes have demonstrated the subjectivity in these intuitive designs. Another possible reason is the current citation sentiment studies have not considered scientific authoring support as a major goal, because they had been strongly influenced by scientometrics and applied linguistics studies.

Nevertheless, to fill in the gap, a common approach is to ask the researchers directly through surveys or interviews. However, as we have seen in the illustrative example, researchers themselves routinely annotate citation sentiments in their literature review process. If their practice is recorded in documents, these documents would be a perfect data set for conducting a non-obtrusive retrospect user study. Luckily such documents do exist – at least in the biomedical domain, researchers have been documenting their manual citation sentiment analysis for detecting citation bias, and the results were published in the form of systematic review. A famous example is (Greenberg, 2009), which was published in BMJ, accompanied by an editorial addressing the citation bias problem (Fergusson, 2009). Greenberg (2009) combined manual citation sentiment analysis and citation network analysis to visualize the evolution of citation bias and the amplification process over time. He examined 218 papers and 675 citations addressing the claim that "*β amyloid, a protein accumulated in the brain in Alzheimer's disease, is produced by and injures skeletal muscle of patients with inclusion body myositis (IBM)*". As a domain expert, Greenberg was aware of empirical evidence against this claim, although this claim had been widely accepted in relevant literature. He identified 10 primary data papers, four of which came from one lab and provided data to support this claim, and the other six reported negative results, including two from the same lab with positive results in earlier studies. These papers are the origins of all citation paths. He then examined the other papers like reviews, annotating the sentiment of each citation statement as "*supportive*", "*neutral*", and "*critical*" toward this claim. In the end only 21 critical citations were found, a stark contrast to 636 supportive citations.

Use case collection

Inspired by Greenberg's work, we searched similar studies in PubMed in order to identify the common use scenarios from biomedical researchers' practices. Because the term "citation sentiment" has not appeared in PubMed metadata, we expanded our query to general citation bias analysis. Query "citation[Title] AND citation[Text Word] AND bias[Text Word]" returned 50 articles. We examined each article to determine whether the study was devoted to citation bias analysis and the individual citation statements were manually examined. We found three matches: (Ravnskov, 1992), (Greenberg, 2009), and (Shrag et al., 2011).

Strictly speaking, Ravnskov (1992) did not examine citation contexts, but instead annotated the polarity of each paper toward a certain claim, and then counted the number of citations to positive and negative claims. We included it to further examine the papers that cited these three papers (59 papers in total). In this round we found three more matches (Cope & Allison, 2004), (Fiorentino et al., 2011), and (Matricciani et al., 2011). In the end we found six relevant papers in 109 papers.

Citation sentiment analysis in biomedical reviews

All six papers belong to the systematic review genre, each paper investigating a specific biomedical claim. Therefore, all papers followed the systematic review methodology to search for a collection of relevant papers as the first step. The relevant papers were then categorized into primary data papers or secondary papers like reviews that cited the primary data papers; each paper's polarity was identified as supporting or opposing the claim. This common behavior indicates that biomedical researchers gave more weight to primary data papers than others. Earliest primary data papers are also the roots of all citation paths.

After the above steps the six studies differed in some ways in the process of examining citation sentiment. Here we group them into three types. Type I did not actually annotate citation sentiment (Ravnskov, 1992; Fiorentino et al., 2011). After annotating each paper's conclusion as positive and negative to the claim, the authors just used the numbers of citations to positive and negative claims to assess citation bias. This approach assumed all citations to positive claims are also positive, and thus did not distinguish two types of opinions: (1) a disagreement with the cited positive claim, and (2) agreement with the cited negative claim (see Figure 1 for illustration). See below an example of the first type:

> "*In contrast to the findings of Sarkozi et al[32] we were unable to show an increase in mRNA for βPP in IBM fibers by using antisense RNA probes.*" (Greenberg, 2009)

Type II annotated each citation statement as supportive, critical, or neutral toward the claim (Greenberg, 2009; Shrag, 2011). Note that the target of their citation sentiment is the claim, not the cited work, while our previous definition targets the cited work. For example, in (Greenberg, 2009) the following citation statement is annotated as *"critical"* to the claim, and it should be annotated as *"supportive"* to the target paper #121, based on the reporting verb *"confirmed"* (Thompson and Ye, 1991).

[#2->#121] *"Microarray studies, such as that done by **Greenberg et al.[52]**, confirmed the increased expression, at the mRNA level, of β-amyloid, ApoE, SOD2, BAX, low-density lipoprotein receptors, and low-density lipoprotein receptor-related protein, but also found all of these genes overexpressed in other inflammatory myopathies, some at much higher-fold ratios than in sIBM."*

Type III examined not only the citation sentiment but also its validity (Cope & Allison, 2004; Matricciani, et al., 2011), indicating researchers' need for citation validity check. The sentiment strength was also examined to make sure the certainty levels were the same in citing and cited papers. For example, if the cited paper used "*suggest*" and the citing paper used "*confirmed*", the citing paper would be overstating. For this problem, Wan et al. (2010) actually conducted a survey and found that researchers often need to check the cited papers to make sure the citing statement did not misrepresent the cited paper. According to researchers' feedback, misrepresentations are common. Wan et al. subsequently developed a web browser enhancement to allow users to conveniently compare a citing statement and its corresponding content in the cited paper.

The following two subsections further discuss two studies, (Ravnskov, 1992) and (Greenberg, 2009), not only because they inspired the other studies, but also because of their unique characteristics.

PubMed's Comment-In-Comment-On function (CICO)

Ravnskov (1992) examined the citation counts of 22 clinical trials to see how the claim that "lowering cholesterol values prevents coronary heart disease" was received in the research community. The citation counts showed that the studies that supported this claim were cited six times more often than the unsupportive ones, while the numbers of supportive and unsupportive trials equaled. Based on this finding, Ravnskov concluded that this claim is not true, although doctors have been advising patients with coronary heart disease to lower cholesterol values in clinical practice.

How is Ravnskov's work received by fellow researchers then? Google Scholar listed more than 300 citations with (Egger et al., 1997) at the top of the list. Egger et al. used Ravnskov's work as an example of the citation bias problem in meta-analysis and received more than 8,000 citations, demonstrating a strong community interest in this problem. But, it is the PubMed's CICO function ("comment-in comment-on") that connects readers to a rebuttal that published as a set of letters to the editor in the same journal. In one letter, Game and Neary (1992) pointed out that Ravnskov's study itself exhibited citation bias, such as excluding a major 11-year supportive trial, including unsupportive early results, etc.

PubMed's CICO function links a research paper to a commentary document (e.g., editorial or letter to editor) that commented the paper. A "comment-in" link that points to the commentary document is embedded in the research paper's web page, and a "comment-on" link pointed to the research paper appears in the commentary document's web page. CICO is probably the first primitive citation sentiment analysis tool, while the citation polarity is not yet categorized, and the current practice is limited to identifying citations in commentary materials only. It started as a manual process, and researchers at the National Library of Medicine have started the research on automating the process recently (Kim et al., 2012).

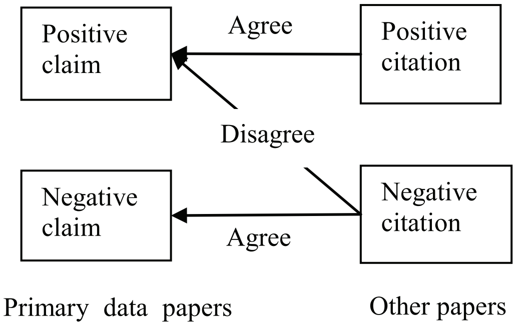

Figure 1. Relationship between citing and cited papers

A closer look at the negative citations in (Greenberg, 2009)

Different from all other studies, Greenberg (2009) published the entire annotation records in a 121-page supplementary document, which provides a valuable opportunity to examine how domain experts annotated individual citations. Given the importance and rareness of negative citations in literature review, we focus on the negative citations only. Because co-citations like [#73->#75, #80, #143] were counted as multiple citations, the 21 negative citations correspond to 17 unique citation statements, 8 with explicit cues and 9 without. To save space we do not include these papers in the reference list, but readers can use these reference numbers to find full references in (Greenberg, 2009).

In these 8 explicitly negative citation statements, half of them reported negative results by combining negations and reporting verbs like *“were unable to show”*, *“less ... than has been reported by others”*, *“could not demonstrate”*, and *“raise the question of”*. The other half stated a negative claim that the protein is not specific for IBM. These linguistic cues may be picked up by computers to infer the citation sentiments.

[#73->#75] *“In contrast to the findings of* ***Sarkozi et al[32] we were unable to show*** *an increase in mRNA for βPP in IBM fibers by using antisense RNA probes.”*

[#73->#75,#80,#143] *“We found considerably* ***less*** *β-amyloid* ***than has been reported by others[6,28,29]****.”*

[#14->#178] *“Furthermore, the intracellular accumulation of amyloid-related proteins, β−amyloid precursor protein (β-APP), phosphorylated tau, presenilin 1, apolipoprotein E, and oxidative stress proteins are also observed in other conditions, leading to the conclusion that* ***they may not be specific to*** *the vacuoles of sporadic IBM[47]”*

[#45->#121] *“The expression of genes that result in the intracellular accumulation of ABPP, tau,…***is not unique to** *IBM, because these genes are equally expressed in sporadic IBM, polymyositis, hereditary IBM, and other myopathies* ***[48]****…”*

[#54->#136] *“Vacuolated fibers with the same deposits, including amyloid,* ***are not specific for*** *IBM; they have been seen in several other chronic distal myopathies such as yofibrillar, desmin, and even in chronic neurogenetic disorders, such as old paralytic poliomyelitis* ***[13]****”*

[#121->#136] *“It has been noted that muscle specimens from patients with postpoliomyelitis syndrome and chronic, long-standing neurogenic weakness have vacuolated muscle fibers with 15 nm filaments immunoreactive for β-amyloid and ubiquitin in a pattern identical to IBM, suggesting that* ***the findings are not specific for*** *IBM****[58]****”*

[#160->#72] *“Our results, and the observations that APP mRNA and protein levels are increased in the developing neuromuscular junction and in regenerating muscles in a variety of neuromuscular and muscle diseases,****[27] raise the question of*** *the significance of the elevated levels of APP in muscle.”*

[#274->#71] “***Sherriff et al. could not demonstrate*** *β-amyloid protein, tau, apoE, or prion protein immunoreactivity in either frozen or paraffin sections of muscle from patients with IBM, despite the use of antigen retrieval technique****[27]****”*

In contrast, the other 9 citation statement did not leave any explicit linguistic cues: 4 of them used expressions that significantly deviated from the original claim, and thus require domain knowledge to infer the citation sentiment; the other 5 used alternative expressions like *“x is also found in z”* or at least *“also”* to re-write the same negative claim (*x is not specific to y*) in a neutral tone. It would be a great challenge for computers to automate this reasoning process. Some linguistic resources like antonyms and synonyms may be helpful for this task.

[#2->#121] *“Microarray studies, such as that done by* ***Greenberg et al.[52]****, confirmed the increased expression, at the mRNA level, of β-amyloid, ApoE, SOD2, BAX, low-density lipoprotein receptors, and low-density lipoprotein receptor-related protein, but* ***also found*** *all of these genes overexpressed in* ***other*** *inflammatory myopathies, some at much higher-fold ratios than in sIBM.”*

[#34->#78] *“Recent observations regarding the interaction of T cells with muscle fibres in polymyositis* ***[48]*** *may also have some relevance to inclusion body myositis. The points of membrane interaction of T cells with the invaded muscle fibres were shown to stain most intensely for APP (T cell) and NCAM-1 (muscle fibres) suggesting a possible interaction between these molecules during muscle fibre invasion.*”

[#38->#72] “*…The accumulation of βAPP mRNA is* ***also*** *increased in these regenerating fibers [54,****55]*** *and βAPP mRNA is present in human myotubes in tissue culture, where it becomes downregulated during their development* ***[55]****”*

[#47->#121] *“An important gene expression profiling study has also found increased expression of amyloid-[beta] and ApoE in IBM, but significantly elevated levels of the same genes were* ***also*** *demonstrated in PM and DM, suggesting that accumulation of such proteins in IBM may be due to posttranscriptional events* ***[43]****.”*

[#71->#80, #143] *“The discrepancies in Aβ immunostaining are difficult to explain, particularly as one of the antisera used in this study (R1280) were also used in previous reports* ***[4,11]****”*
[#89->#100] *“Alternatively, Aβ-intracellular deposition may be an epiphenomenon unrelated to myofiber death* ***[Pruitt et al. 1996]”***

[#90->#70] *“it has been shown that human macrophages found in muscle of various muscle diseases, including IBM, demonstrate strong immunoreactivity and mRNA for β-amyloid precursor protein (βAPP), suggesting that the abnormal accumulation of the correspondent protein is generated, at least partly, by locally increased transcription outside vacuolated muscle fibers****[23]****.”*

[#106->#72] *“Because Nogo-A is increased in regenerating muscle fibers [42], it may have additional roles in those young fibers. One possibility might be to help manage the increased AβPP known to occur in regenerating muscle fibers in vivo [4,* ***29]****”*

[#121->#70] *“β-Amyloid mRNA is* ***also found*** *within macrophages in a variety of muscle diseases[59].”*

[#144->#70, #75] *“Immunohistochemical studies, however, have found overexpression of βAPP transcripts not just in a small percentage of abnormal fibers in IBM, but* ***also*** *in regenerating muscle fibers in various* ***other*** *muscle diseases* ***[27,28]****.”*

What can we learn from biomedical researchers?

The above six studies have not been cited by any automated citation sentiment studies, suggesting a disconnection between the two fields. We now compare the biomedical researchers’ approaches and current automated approaches, and discuss how to improve the automated design for better usability.

No need for fine-grained citation function classification

In literature review process biomedical researchers focused on the sentiment of specific claims and their citations. Different from research administrators, the researchers did not conduct fine-grained assessment of the cited papers' research contributions. For example, perfunctory and organic citations were not distinguished. Instead, this task is avoided by creating a body of relevant literature through systematic search in bibliographic databases like PubMed. All included papers are deemed relevant to the claim and thus every citation is worth reading. Even if the user is not conducting systematic review, perfunctory and organic citations can be distinguished by examining the relevance between the full texts of the citing and cited papers rather than limiting the analysis in the short citation statement.

Definition of citation sentiment

For researchers, the current polarity based citation sentiment analysis is not precise enough. When discussing a scientific claim, the sentiment target may be the claim or the cited paper. Citation sentiment strength and validity are also important for researchers. Therefore a richer representation would be needed, such as a tuple consisting of the citation source, target, polarity, strength, and validity. This representation seems closer to the definition of fine-grained opinion analysis (Wiebe et al., 2005; Liu, 2012) rather than a simplified definition of polarity based sentiment analysis, which assumes the opinion holder and target are known (Pang & Lee, 2008).

Integrating citation sentiment to biomedical metadata

Biomedical researchers gave more weights to primary data papers than review papers when examining the strength of their conclusions. PubMed has a classification of article types, and the strength of their conclusions varies. These metadata play important roles in researchers' literature search and review. Therefore, citation sentiment studies and biomedical ontology studies should be mutually informative toward future integration. For example, citation sentiment is one element in Shotton's citation typing ontology for biomedical literature (Shotton, 2010). If the two fields develop separately, their definitions may differ and thus affect the inter-operability.

Integrating domain knowledge in citation sentiment identification

Citation sentiment can be very difficult to identify. In Greenberg's study more than half of the negative citation statements did not contain any explicit linguistic cues, making it hard for computers to make decision. The fact that an expert like Greenberg was able to do so indicates the need to integrate domain knowledge into automated citation sentiment analysis, and there are at least two possible ways.

First, utilize research results from relevant bioinformatics research. Many NLP researchers are particularly working on bioNLP applications, and one of them is to automatically classify scientific claims by their polarity and certainty (Light et al., 2004; Medlock, 2009; Battistelli & Amardeilh 2009; Blake, 2010). The polarity of the claim and the polarity of the citation can be compared to determine the citation sentiment.

Second, gather researchers' manual annotations as training data to improve automated systems. The legacy annotations in published studies are valuable, but poor data management may result in data loss (Yu & Ku, 2010). A more viable approach is to build an interactive platform to allow researchers to provide feedback. Given the importance of citation sentiment identification, researchers may be willing to invest their precious time to contribute their feedback to improve the automated systems, just as email users are willing to correct the mistakes that spam filters make (Gormack & Lynam, 2007).

CONCLUSIONS

This study used a non-obtrusive retrospect approach to identify researchers' needs for citation sentiment analysis by reviewing their manual citation sentiment analyses as use cases. We identified a number of differences between their approaches and current automated studies. Instead of targeting the cited papers, researchers were found to set the specific scientific claims as the target of citation sentiment. They also examined more citation sentiment aspects like strength and validity than the simple polarity based sentiment. Researchers focused their analyses on claim and citation sentiment and did not pursue more fine-grained citation function classification proposed in scientometrics studies.

These differences reflected the researchers' unique needs for citation sentiment analysis to obtain comprehensive and reliable opinions from prior literature, and provided insights to improve current automated analysis toward more user-centered design. Although our study analyzed six use cases only, given the similarity of literature review process in different disciplines, it is reasonable to expect similar needs applies to other disciplines as well (Evans & Foster, 2011; Glass & Smith, 1979).

REFERENCES

Angrosh, M.A., Cranefield, S., & Stanger, N. (2010). Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries. *Proceedings of JCDL 2010*, 293-302.

Athar, A. (2011). Sentiment analysis of citations using sentence-structure-based features. *Proceedings of NAACL-HLT 2012*, 597-601, Portland, OR, June 19-24, 2011.

Athar, A. & Teufel, S. (2012). Context-enhanced citation sentiment detection. *Proceedings of NAACL-HLT 2012*, 597-601, Montreal, Canada, June 3-8, 2012.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: a networkanalytical model. *American Sociological Review* 63, 829-846

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Computers and Biomedical Research 43*(2), 173-189.

Bonzi, S. (1982). Characteristics of a literature as predictors of relatedness between cited and citing works. *JASIST 33(4)*, 208-216

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation 64*(1), 45-80.

Case, D. O. & Higgins, G. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *JASIST 51(7)*, 635-645.

Chubin, D. E. & Moitra, S. D. (1975). Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science*, 5(4), 423-441

Cope, M. & Allison, D. (2010) White hat bias: examples of its presence in obesity research and a call for renewed commitment to faithfulness in research reporting. *International Journal of Obesity, 34,* 84-88.

Cronin, B. (1984). *The citation process; the role and significance of citations in scientific communication.* London: Taylor Graham.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ 315*, 629-634.

Evans, J. A., & Foster, J. G. (2011). Metaknowledge. *Science 331*(6018), 721.

Fergusson, D. (2009). Inappropriate referencing in research. *BMJ*, *339*, b2049.

Fiorentino, F., Vasilakis, C. & Treasure, T. (2011) Clinical reports of pulmonary metastasectomy for colorectal cancer: a citation network analysis. *Br J Cancer 104(7)*, 1085-1097.

Game, F. L. & Neary, R. H. (1992). Frequency of citation and outcome of cholesterol lowering trials. *BMJ 305*, 421-422.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics 1(4)*, 359-375

Garzone, M. & Mercer, R.E. (2000). Towards an automated citation classifier. *Advances in Artificial Intelligence, Lecture Notes in Computer Science 1822*, 337-346.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational evaluation and policy analysis 1*(1), 2-16.

Cormack, G. V., & Lynam, T. R. (2007). Online supervised spam filter evaluation. *ACM Transactions on Information Systems (TOIS) 25*(3), 11.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ 339*, b2680.

Hayes, P. J., Andersen, P. M., Nirenburg, I. B., & Schmandt, L. M. (1990). TCS: A shell for content based text categorization. *Proceedings of the IEEE Conference on Artificial Intelligence Applications*, 1990.

Herring, S. C. & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4), 439-459.

Hodges, T. L. (1972). Citation indexing: its potential for bibliographical control. *Ph.D. thesis*, University of California at Berkeley.

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes 13(3)*, 239–256.

Hyland, K. (1999). Disciplinary discourses: Writer Stance in Research Articles', in *C. Candlin and K. Hyland (eds) Writing: Texts: Processes and Practices*, 99–121.London: Longman

Kaplan, D., Iida, R., & Tokunaga, T. (2009). Automatic extraction of citation contexts for research paper summarization: a co-reference-chain based approach. *Proceedings of IJCNLP 2009*, 88-95

Kim, I. C., Le, D. X., & Thoma, G. R. (2012). Identifying "comment-on" citation data in online biomedical articles using SVM-based text summarization technique. *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI'12)*, Las Vegas, Nevada, July 16-19, 2012.

Lipetz, B-A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16(2), 81-90.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1-167.

MacRoberts, M.H. & MacRoberts, B.R. (1984). The negational reference: or the art of dissembling. *Social Studies of Science*, 14(1), 91-94.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *JASIS 40*(5), 342-349.

Matricciani, L, Olds, T & Williams, M. (2011). A review of evidence for the claim that children are sleeping less than in the past. *Sleep, 34(5)*, 651-659.

McCain, K. W. & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics 17(1-2)*, 127-163.

Mercer, R. E. & DiMarco, C. (2004). A design methodology for a biomedical literature indexing tool using the rhetoric of science. *Proceedings of HLT-NAACL 2004 Workshop on Linking Biological*

Literature, Ontologies and Databases (BioLINK 2004), 77–84, Boston, Massachusetts, USA.

Moravcsik, M.J. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5(1), 86-92.

Nanba, H., Kando, N., & Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Proceedings of 11th SIG/CR Workshop*, 117–134.

Peritz, B.C. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics*, 5, 303-312

O'Connor, J. (1982). Citing statements: computer recognition and use to improve retrieval. *Information Processing and Management*. 18(3), 125-131.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval,* 2(1-2):1-135.

Piao, S., Ananiadou, S., Tsuruoka, Y., Sasaki, Y., & McNaught, J. (2006). Mining opinion polarity relations of citations. *Proceedings of the International Workshop on Computational Semantics (IWCS)*, 366–371.

Qazvinian, V. & Radev, D. (2008). Scientific paper summarization using citation summary networks. *Proceedings of COLING 2008,* 689-696.

Ravnskov, U. (1992). Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ 305*, 15-19.

Ritchie, A., Robertson, S. & Simone Teufel. (2008). Comparing citation contexts for information retrieval. *Proceedings of CIKM 2008*, 213–222.

Schafer, U. & Kasterka, U. (2010). Scientific authoring support: a tool to navigate in typed citation graphs. *Proceedings of NAACL-HLT 2010 Workshop on Computational Linguistics and Writing*, 7-14.

Schafer, U., & Spurk, C. (2010). TAKE scientist's workbench: semantic search and citation-based visual navigation in scholar papers. *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing (ICSC),* 317-324.

Schrag, M., Mueller, C., Oyoyo, U., Smith, M. A., & Kirsch, W. M. (2011). Iron, zinc and copper in the Alzheimer's disease brain: a quantitative meta-analysis. Some insight on the influence of citation bias on scientific opinion. *Progress in Neurobiology 94(3)*, 296-306.

Schwartz, A. S., Divoli, A., & Hearst, M. A. (2007). Multiple alignment of citation sentences with conditional random fields and posterior decoding. *Proceedings of EMNLP-CoNLL 2007,* 847–857, Prague, June 2007

Shotton, D. (2010). Cito, the citation typing ontology. *Journal of Biomedical Semantics*, *1(Suppl 1),* S6.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics 87*, 373-388.

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics 7(1)*, 39-56

Swales, J. (1990). *Genre Analysis: English in Academic and Research Setting*. Cambridge University Press

Teufel, S. Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function. *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 80-87, Sydney, July 2006.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of EMNLP 2006*, 103-110

Thompson, G. & Ye, Yiyun. (1991). Evaluating in the reporting verbs used in academic papers. *Applied Linguistics 12(4)*, 365-382.

Wan, S., Paris, C. & Dale, R. (2010). Supporting browsing-specific information needs: introducing the citation-sensitive in-browser summarizer. *Web Semantics: Science, Services, and Agents on the World Wide Web, 8,* 196-202.

White, H. D. (2004). Citation analysis and discourse analysis revisited. *Applied Linguistics, 25(1)*, 89-116.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3), 164-210.

Yu, B. and Ku, M. (2010). Collecting legacy corpora from social science research for text mining evaluation. *Proceedings of ASIST 2010 Annual Meeting*, Pittsburgh, PA, October 22-27, 2010

Zhang, X., Qu, Y., Lee Giles, C., & Song, P. (2008). CiteSense: supporting sensemaking of research literature. *CHI'08*, 677-680.