Syracuse University
Citation Opinion Retrieval and Analysis (CORA):
An Automated Plug-in Tool for Digital Libraries

Abstract

Academic libraries are facing demands for more sophisticated services such as retrieving relevant literature and evaluating researchers' academic contributions. However, current digital libraries have not been able to provide relevant functions to support these services. For example, given an article, common bibliographic tools like PubMed and Google Scholar have provided citation counts and URL links to citing works; however, navigating through the large number of citations is still a daunting task for researchers and librarians. A significant amount of time is required for reading the citing works to understand their opinions toward the cited works, which is particularly challenging for inexperienced student researchers, or researchers who are entering a new or interdisciplinary field. Without effective citation content analysis tools, citation bias and inaccurate citations may remain undiscovered for many years, which also affect the reliability of citation count-based assessment of research impact.

Since the 1960s, many studies have attempted to identify the citation statement and categorize citation functions. However, automated citation opinion analysis was not explored until recently, when pilot studies, equipped with machine learning and natural language processing techniques, emerged to explore computational approaches to identify citation opinions. These pilot studies marked citation opinion analysis as a new research area, and also acknowledged the great challenges toward building an accurate citation opinion analysis tool due to the unique characteristics of scientific criticism.

This early career project will be a substantial effort toward building an automated tool that can plug into a full-text bibliographic database, extract the citation statements toward a cited article, separate substantial citations from perfunctory ones, and categorize substantial citation opinions by their purposes (e.g. comparison, critique, etc.), subjects (e.g. methods, results, etc.), and tones (e.g. positive, negative, and neutral). This Citation Opinion Retrieval and Analysis tool, abbreviated as CORA, will save librarians and researchers significant amount of time to find the most useful comments from a large number of citations. CORA will also provide a new, qualitative approach for assessing research impact. CORA can also help monitoring the quality of scientific publications by facilitating easier identification of citation bias and inaccurate citations from the re-organized citations.

Syracuse University's Dr. Bei Yu is the PI applying for this early career grant. No other senior personnel will participate in this project. This project spans over a three-year period, starting from June 1, 2014. In Year 1 an annotated corpus consisting of research articles from two scientific disciplines (biomedicine and natural language processing) will be developed to train and evaluate CORA. In Year 2 a baseline system will be developed based on previous studies on citation behavior and linguistic characteristics of scientific criticism, and then a number of machine learning and natural language processing techniques will be applied to improve the baseline system. Both the baseline and the improved systems will be evaluated on the annotated corpus (the gold standard) to examine the performance improvement. In Year 3 CORA will be plugged into a website that simulates two real digital libraries (PubMed Central Open Access Subset and the ACL Anthology). User feedback will be collected to evaluate CORA's accuracy in real-world scenario, and usefulness in assisting researchers in scholarly work. An annual workshop will be organized to invite librarians and researchers on campus and from nearby institutions to focus-group discussions on the system performance, website usability design, and potential use scenarios for CORA.

CORA is expected to benefit librarians and researchers in all scientific disciplines, and may be extended to the humanities and non-English publications in the future. The annotated corpus and CORA source code will be publicly available to encourage further research in this area, in order to create an open source community to support continued citation opinion analysis software and extension of the training corpus. Collaboration with large digital libraries such as PubMed and Google Scholar will be sought for beta test in order to release CORA as a real plug-in tool for academic digital libraries.

Syracuse University
Citation Opinion Retrieval and Analysis (CORA):
An Automated Plug-in Tool for Digital Libraries

1. Statement of need

Academic libraries are facing demands for more sophisticated services such as retrieving relevant literature and evaluating researchers’ academic contributions. However, current digital libraries have not been able to provide relevant functions to support these services. For example, given an article, common bibliographic tools like PubMed and Google Scholar have provided citation counts and URL links to citing works; however, navigating through the large number of citations is still a daunting task for researchers and librarians. This early career project aims to build an automated tool that can retrieve and extract citation statements from articles that cite a given article, separate substantial citations from perfunctory ones, and categorize substantial citation opinions by their purposes (e.g. comparison, critique, etc.), subjects (e.g. methods, results, etc.), and tones (e.g. positive, negative, and neutral). This citation opinion retrieval and analysis tool, abbreviated as CORA, will address five serious problems existing in current scholarly communication.

<u>Problem 1</u>: *Researchers spend a lot of time gathering citation opinions; it is rarely done comprehensively*.

An example best illustrates how researchers routinely extract and categorize citation opinions during literature review, and how this routine operation is extremely time-consuming. Assume a researcher comes across an article (Hayes et al., 1990), which presents a knowledge-based text categorization system CONSTRUE. This article seems quite influential at first because it had been cited 113 times in Google Scholar as of September 12th, 2013. These citations are further examined to see what other researchers thought of the CONSTRUE system. Google Scholar ranks the citations by the citing articles’ own citation counts. The opinions in the top five citations are manually extracted and annotated (Table 1): citation #1 is explicitly negative, questioning the result’s reliability and the validity of the test data; #2 mixed positive and mitigated negative opinions on the test data’s appropriateness; #3 seems literally neutral, however, because it used CONSTRUE as the benchmark system, it is reasonable to infer that #3 is implicitly positive; #4 is explicitly positive, and #5 made a neutral statement without detailing its relationship to the cited article. The opinions in the top five citations have shown some concerns on CONSTRUE’s effectiveness, but a thorough scan of all citations is necessary to find out whether other concerns have also been raised, or whether earlier concerns have been further addressed in later publications.

Assuming a researcher uses five minutes to examine one citation, including the activities of opening the full-text PDF files, finding the citation statements, and understanding the polarities of the opinions. With the same speed it would take 9 hours to examine all citations to this article. In reality, researchers can only afford to examine a small portion of the citations, resulting in an incomplete view of the citation landscape.

	Opinion	Citation statement
1	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.”_
2	Mitigated negative	_“Recent work has shown that in certain environments, knowledge-based systems can do code assignment **quickly and accurately** [Hayes and Weistein 1991; **Hayes et al. 1990**]… 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. ”_
3	Implicit 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].”**_
4	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].”_
5	Neutral	_“The training documents can be used by human experts to generate categorization rules ([1], **[7]**) or …”_

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

Problem 2: Incomplete citation scan particularly jeopardizes interdisciplinary research.

An incomplete citation scan particularly harms researchers who are entering new fields, especially interdisciplinary fields, not only because they may not have acquired adequate domain knowledge to grasp the whole picture, but also because of disciplinary differences in scientific criticism. Hyland (1999) found that "hard science" (e.g., biology, physics) writers use significantly fewer negative critiques than "soft science" (e.g. social sciences such as sociology, marketing) writers did, probably because different opinions are more likely to co-exist in social science and humanities publications, due to the extreme complexity in social phenomena. With the increasingly popular data-driven scholarship and computational social science research (Lazer et al, 2009), more computational scientists are engaged in social science research problems. It is very important for them to browse all existing opinions, especially the critical ones, to guide their cross-disciplinary research.

The following example best illustrates this problem. Linguists, psychologists, and communication scholars have all studied a research question "do men and women use language differently?" On the one hand, many studies have found differential patterns (e.g. Lakoff, 1975; Herring, 1992; Holmes, 1993; Biber et al., 1998; Coates & Johnson, 2001; Koppel et al., 2003; Newman et al., 2008; Pennebaker, 2011). On the other hand, a number of studies claim that these differences are merely the artifact of the communication context (e.g. Rubin & Greene, 1992; Krauss & Chiu, 1998; Janssen & Murachver, 2004; Herring & Paolillo, 2006). Communication Accommodate Theory (CAT) further suggests that people may adapt their language use styles to converge with their communicative partners' styles to gain social approval, resulting in less pronounced gender difference in mixed-gender communications (Giles & Coupland, 1991; Yu, 2011; 2013a). These mixed results indicate a complicated relationship between gender and communication context in language use.

With the development of Computational Linguistics, computer scientists joined in the research effort. Their main interest has been in building automated tools to identify author gender (e.g., de Vel et al., 2002; Koppel et al., 2003; Yan & Yan, 2006; Mukherjee & Liu, 2010). These studies often assume gendered language as an established fact and rarely cite the different opinions articulated in the social science literature. Consequently, communication context is neglected in some studies, resulting in less than ideal research outcomes. For example, de Vel et al. (2002) aimed to identify the gender of email senders, but did not consider the possibility that an email receiver's gender may affect the sender's language use for communication accommodation purpose. If computer scientists were equipped with a citation opinion analysis tool like CORA, they would be able to quickly identify both positive and negative results in language and gender studies, and thus take both into consideration in their research design.

Problem 3: Citation biases threaten research validity, but are difficult to monitor.

Citation bias refers to the phenomenon that negative results, including non-significant findings and discordant opinions, receive many fewer citations than positive results, leading the research community to a distorted view of current research status (Greenberg, 2009). Citation bias may have occurred in the aforementioned research on gender difference in language use, but the very task of identifying citation bias requires thorough citation opinion analysis, which, has only been manually conducted after a domain expert and whistleblower initiated an investigation (e.g., Fergusson, 2009; Greenberg, 2009; Fiorentino et al., 2011; Schrag et al., 2011).

Such manual analysis is time-consuming, and can be error-prone if not thoroughly conducted. As an unsuccessful example, Ravnskov (1992) claimed to have found citation bias toward the claim that "*lowering cholesterol values prevents coronary heart disease*", reporting that the supportive studies were cited six times more often than the unsupportive ones, although their numbers were the same. How is Ravnskov's work received by fellow researchers then? Google Scholar lists more than 300 citations; but it is the PubMed's CICO function ("comment-in comment-on") that connects readers to a rebuttal letter, which points out that Ravnskov's study itself exhibits citation bias, as it excluded a major 11-year supportive trial and included unsupportive early results (Game & Neary, 1992).

PubMed's CICO function provides two-way linkages between research papers and their commentaries, letters, editorials, and correspondences (Kim et al., 2012). However, the number of publications in such opinion-rich genres is very small, and citation opinions expressed in the large number of research articles should be examined as the main data source for preventing and detecting citation bias. Manual analysis is not only time-consuming and error-prone, but is also too late to be preventive, since citation bias can only be detected after it is formed. CORA is expected to identify critical citations and citations to negative results, which will be a crucial component for developing an early warning system to monitor citation bias in the future.

Problem 4: Inaccurate citations mislead researchers and the general public.

An inaccurate citation occurs when the citer paraphrases or summarizes the cited work in an inaccurate way, such as, a negative result is cited as positive, a suggestive result is cited as a confirmed finding, etc. (Greenberg, 2009). A survey found that researchers perceive inaccurate citations to be a common phenomenon, and often check the cited articles themselves to verify the accuracy of citing statements (Wan et al., 2010). Inaccurate citations also appear in science news for the general public. For example, the news media has under-reported the risks, and thus exaggerated the potential benefits of neuroscience innovations and prescriptive medication (Moynihan et al., 2000; Cassels et al., 2003; Partridge et al., 2011). A fully-automated citation verification tool is not only useful for researchers in literature review, but also useful for editors and reviewers in the peer review process to ensure accurate citations before papers are published.

Because this project is only 3 years, fully-automated citation verification using CORA is beyond the project scope. However, by project end, CORA is expected to extract and categorize citation opinions, readers will then be able to review the positive and negative citations more conveniently and catch the inaccurate ones more efficiently. In addition, by deepening the knowledge on automated identification of citation polarity and certainty, this project will lay a solid foundation for future work on fully-automated citation verification.

Problem 5: Citation opinion-based measures are needed for qualitative evaluation of research impact.

Measuring research impact is an important component in research administrative decisions, such as tenure and promotion cases, and institutional and government funding policies. However, current approaches heavily rely on quantitative measures based on citation counts, such as the *h*-index. Without considering the citation contexts, these measures suffer from a number of shortcomings, such as bias toward fashionable research topics, and indiscriminate treatment of substantial vs. perfunctory citations and positive vs. negative citations. For example, hundreds of retracted papers were still cited as valid research years after retraction (Kochan & Budd, 1992; Budd et al., 1998). The less media coverage a retracted paper receives, the more likely it continues to be cited (Unger & Couzin, 2006). A tool like CORA is expected to distinguish the accurate, negative citations that mention retractions from the inaccurate, positive ones, and also push the retractions to readers' immediate attention.

CORA is expected to provide an alternative, qualitative approach for assessing research impact, which may provide new evidence to resolve controversies in current scientometrics studies. For example, the hypothesis that female researchers produce fewer but higher quality publications has been tested with mixed results based on counting the number of publications and citations (Long, 1992; Symonds et al., 2006). Citation opinion analysis is expected to provide a different measure of research quality and impact, and thus provide new evidence for understanding the phenomenon of gender inequality in STEM research.

2. Impact

CORA will add new, important functionality to full-text bibliographical databases to assist librarians and researchers with efficient and effective navigation through numerous citations, which will further enable fast and comprehensive literature review, better opportunity to detect citation bias and inaccurate citations, and content-based evaluation of research impact. To reach this goal, the PI has set the following project objectives with measurable outcomes: (1) design a cross-disciplinary citation opinion classification scheme that serves as the framework to categorize citation opinions; (2) design and evaluate automated approaches for classifying citation opinions; (3) build a prototype website to simulate a digital library environment and conduct a large-scale user study to test CORA's usefulness in assisting researchers with scholarly work.

The CORA project will advance the knowledge of automated recognition of opinion expression in scientific literature, and transform the current bibliographic databases to next-generation "smart" systems equipped with automated citation opinion retrieval and analysis functions. This project focuses on building accurate citation opinion extraction and classification algorithms. The longer-term goal is to integrate a highly-accurate CORA system into real full-text bibliographical databases, such as PubMed and Google Scholar, and thus provide a complete solution to address the aforementioned problems in scholarly communication.

Besides contributing to the research community, this research is also expected to broadly benefit society in the following ways:

- Given the increasing impact of scientific discoveries on people's everyday life and government policies, such as evidence-based medicine, CORA will help monitor the validity of scientific research, ensuring reliable evidence for supporting individual and government decision making in healthcare, education, and many other important aspects of life.
- Research impact assessment provides important evidence for scientific policy making. CORA's qualitative approach for citation opinion analysis will complement current quantitative approaches toward more accurate assessment.
- The algorithms developed in this project can be further applied to analyzing the citation opinions in scientific news, comparing them against the actual scientific findings, thus monitoring the quality of scientific communication with the general public.
- In addition to assisting research, CORA may also be used as an educational tool to facilitate classroom discussions on given research topics. Teachers and students can quickly locate substantial positive and negative opinions regarding a research topic and focus their discussions on the important citations.
- This project's evaluation plan includes an annual CORA user workshop, which will raise the awareness of the citation problems and promote appropriate citation behavior among faculty and students at Syracuse University and nearby institutions.

3. Project design

This project's research approach draws on prior studies on sentiment classification, opinion mining, and citation analysis from various disciplines such as bibliometrics, academic writing, and legal information retrieval. This project will combine prior knowledge on the linguistic characteristics of scientific criticism, machine learning, and natural language processing techniques to tackle the citation opinion analysis problem.

3.1. Choosing the work domains

Due to disciplinary differences in scientific discourse and opinion expression (Hyland, 1999), it remains an open question whether a citation opinion analysis method trained on citations in one domain (e.g. biomedicine) can be effectively applied to another domain (e.g. natural language processing, a.k.a NLP). This project chooses biomedicine and NLP as two parallel work domains. The research will be carried out on data sets from both domains, and evaluate the algorithms' domain transferability.

These two disciplines are chosen for their researchers' strong interests in citation opinion analysis (Greenberg, 2009; Abu-Jbara et al., 2013; Yu, 2013b) and the availability of large, open-access, and citation-indexed corpora. PubMed Central (PMC) is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature. Because PMC does not allow web crawlers, this project will use the PMC Open Access Subset (PMC-OA), a small part of the PMC collection available for FTP download under the Creative Commons license. The current version contains nearly 600,000 articles. The ACL Anthology Network Corpus (AAN) is the citation-indexed version of the ACL Anthology, the major full-text bibliographic database in NLP. The current version of AAN contains nearly 20,000 articles. This project focuses on biomedical and NLP articles written in the English language. Future research would extend CORA to publications in multiple languages and in social sciences and humanities.

3.2. Identifying citation statement boundaries

Identifying citation statement boundary is not a trivial task: a citation statement may include a few sentences before and after the citing sentence; multiple citation statements may even overlap (see Table 1 for an example). To identify the citation statement boundary, we will first implement a rule-based algorithm as baseline, and then improve it with machine learning approaches. Rule-based approaches (e.g., O'Connor, 1982; Nanba et al., 2000; 2004) utilize the cue words or phrases (usually transition words like "however" and pronouns like "these") that connect the sentences within a statement. Machine learning approaches treat this problem as a sequential labeling problem and then apply common structured classification methods (Abu-Jbara et al., 2013). Co-reference resolution techniques (Kaplan et al., 2009) may be useful for finding "nicknames" that refer to the same opinion objects and to further improve accuracy. The above approaches have been tested effective in small annotated data sets in individual domains. They will now be tested in a new, large annotated corpus, and performance in two different domains will be compared.

3.3. Defining the citation opinion annotation scheme

To automate citation opinion classification, the first step is to create a classification scheme. Historically, bibliometrics studies have attempted to classify citations by their "functions", specifically, when *x* cites *y*, whether *x* is questioning *y*'s findings, using *y*'s method, or simply paying homage to *y*, etc. A number of classification schemes have been proposed since the 1960s (e.g., Lipetz, 1965; Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975; Peritz, 1983; Bonzi, 1982). Developed on small data sets, these schemes lack consensus in defining the categories (Baldi, 1998), and inter-coder agreements were not reported. In recent years, NLP researchers have begun to tackle the citation classification problem (e.g., Garzone & Mercer, 2000; Nanba et al., 2004; Piao et al., 2006; Taboada, et al., 2006; Teufel et al., 2006b; Qazvinian & Radev, 2008; Zhang et al., 2008; Ritchie et al., 2008; Angrosh et al., 2010; Schafer & Kasterka, 2010; Athar, 2011; Athar & Teufel, 2012; Abu-Jbara et al., 2013). To overcome the shortcomings of prior schemes, new schemes were proposed and tested on NLP articles, such as the simplified citation function annotations (Teufel et al., 2006a), citation sentiment annotations (Athar, 2011), and citation purpose and polarity annotations (Abu-Jbara et al., 2013). However, none of the above annotation schemes was particularly tailored toward solving the five scholarly communication problems described in the statement of need.

Inspired from the fine-grained opinion annotation in customer reviews (Hu & Liu, 2004), news articles (Wiebe et al., 2005), and the PI's pilot studies (Yu, 2013b; Yu & Li, 2013), we constructed a multi-dimensional scheme that annotates each citation from four aspects: *purpose, subject, tone*, and *objects* (see Table 1 for an annotation example).

Purpose: Citation purpose is defined as the intention behind a citer's decision to cite a certain reference. We found three common kinds in the "discussion" sections of biomedical articles (the section with richest citation opinions): *comparison, critique,* and *info.* New types may be added as we continue to annotate citations in the "introduction", "methods", and "results" sections. For critiques, we will distinguish explicit and implicit critiques. A citation statement that does not belong to *comparison* or *critique* will be annotated as *info.*

Subject: *Subject* refers to the aspect of the research that was discussed in the citation. We found *method* and *result* to be the main types of subjects in biomedical articles. More types of subject may be added if finer differentiation is necessary, e.g., research goals, hypotheses, claims, etc. All other undefined subjects are assigned to the category "*general*". A citation that is annotated as "purpose=*info*" and "subject=*general*" is usually a perfunctory citation, e.g. "*Sentiment analysis or opinion mining has been an active research area in recent years ([x])*".

Tone: *Tone* is defined as an extension to the concept of polarity. The traditional three-value polarity definition (positive, negative, or neutral) does not correspond well with the *comparisons* because both favorable comparison (e.g. "*x is better than y*") and agreement (e.g. "*x's result is in line with y*'s") would be positive.

Therefore, we added more values, such as *accordance* and *discordance* for result comparisons. To support citation bias monitoring, we also annotate the polarities of citer-paraphrased results: *positive* results are defined as significant findings (e.g. "*strong correlation was found in [x]*"), and *negative* results mean non-significant findings (e.g. "*No improvement was found in [x]*") or discordant results (e.g. "*We could not replicate the result in [x]*"). The results of descriptive studies will be coded as *neutral* (e.g. "*[x] observed 20% decline in …*").

Objects: *Objects* are the entities involved in the citation statements. They are a list of cited articles for *comparisons*, or owners and targets of *critiques*, or *sources* of *info*. Identifying the *objects* is necessary for annotating complicated citation opinions that involve multiple cited work , e.g., "*[x]'s criticism on [y] is unfounded*" contains two negative critiques: one from [*x*] to [*y*] and the other from the citer to [*x*].

<citation source= PMC3328846 >
<A1 purpose=critique subject=method tone=negative owner=[16][17] target="isolated pharmacist-centred interventions">Since the start of our trial, important studies have questioned the effectiveness of isolated pharmacist-centred interventions in general practice. **<A2 purpose=info subject=result tone=negative owner=[17]>**For example, the HOMER trial[17] aimed to assess whether home-based medication review by pharmacists in older people would affect hospital readmission rates. The researchers reported an increase in hospital admissions and no improvement in quality of life or death rate. **</A2><A3 purpose=info subject=result tone=negative owner=[16]>** The RESPECT trial [16] showed no benefit of involvement of community pharmacists in the moderation of drug management (pharmaceutical care) in older people in general practice.**</A3></A1>**

Table 2. An example of citation opinion annotations

3.4. Annotating the training corpus

The above annotation scheme has been tested on a small sample of biomedical articles (~50 articles) with satisfactory levels of intra-coder agreement and inter-coder agreement (Yu & Li, 2013), and will be further improved and then used to annotate 500 PMC-OA articles and 500 AAN articles as training data at the beginning of this project. Annotators will use GATE (Cunningham et al., 2002) to annotate the corpus. GATE has been used by other opinion annotation projects (Wiebe et al., 2005; Somasundaran & Wiebe, 2010). Each article will be annotated by two coders, and inter-coder agreement will be measured for annotation reliability. For citation statement boundary annotation, we use the text span agreement measure developed by Wiebe et al. (2005). For citation purpose, subject, tone, and objects annotation, we will use common categorical agreement measures like Kripendorff's alpha (Passonneau, 2004) and Cohen's Kappa (Wiebe et al., 2003; Stoyanov et al., 2006).

3.5. Multi-dimensional citation opinion classification

The multi-dimensional scheme enables a divide-and-conquer approach that can integrate a number of existing techniques for citation opinion analysis, namely, topic-oriented text classification methods to identify citation *purpose* and *subject* (e.g. Sebastiani, 2002), sentiment classification methods to identify *tones* (e.g. Pang & Lee, 2008), and information extraction and co-reference resolution methods to identify *objects* (e.g. Ng & Cardie, 2002). High accuracy is reasonably expected in citation *purpose* and *subject* classification in that topic-oriented classification methods have been well developed in recent years. The PI will focus on developing new methods for predicting the tones and the relationships among objects.

Classifying citation opinion tones: Although the citation tones are considered very difficult to classify because of their varied nuances and gradations (Peritz, 1983), the satisfactory level of inter-coder agreement in our pilot study (Yu & Li, 2013) and related studies (Athar, 2011; Abu-Jbara et al., 2013) suggests that explicit patterns do exist. Actually, legal citation research services like *KeyCite* and *Shepard's* have been routinely indexing and categorizing citation opinions to case law documents using "editorial phrases" (Taylor, 2000). Similar evaluative phrases and reporting verbs also exist in scientific literature (Thompson & Ye, 1991; Hyland, 1999; Teufel et al., 2006b; Garzone & Mercer, 2000). We will consult prior knowledge in applied linguistics and academic writing and apply a knowledge-based approach to gather linguistic cues from scientific literature.

Such knowledge-based approaches usually result in high precision and low recall due to variations in opinion expressions. To enrich the collection of linguistic cues, we will use the matured supervised learning and feature selection methods to automatically discover more citation opinion indicators from the training corpus, and select the most effective ones for citation opinion classification (Sebastiani, 2002; Hatzivassiloglou & McKeown, 1997; Wiebe et al., 2004; Kanayama & Nasukawa, 2006). Furthermore, compared to the training corpus with 1,000 manually-labeled articles, the entire PMC-OA and AAN corpora contain ~620,000 unlabeled articles. Unsupervised learning approaches can utilize a large amount of unlabeled data to search for more linguistic cues that frequently co-occur with the cues that have already been found using knowledge-based and supervised learning approaches. Some unsupervised learning approaches have been tested effective in customer review analysis, such as the Pointwise Mutual Information (Turney & Littman, 2003), the WordNet synonym relationship (Kamps et al., 2004), the Spin Model (Takamura et al., 2005), and double propagation (Qiu et al., 2011). In this project we will examine their effectiveness in finding linguistic cues as citation opinion indicators.

Despite the existence of explicit linguistic cues, scientific opinions bear some unique characteristics which make citation tone identification difficult. First, unlike opinion-rich documents like customer reviews and political debates, negative citations are rare in scientific publications, and even scarcer in hard science publications (Hyland, 1999), causing difficulty in gathering training examples. Additional machine learning strategies will be used to deal with the problem of small training set, such as active learning (Tong & Koller, 2002) and co-training approaches (Blum & Mitchell, 1998; Pierce & Cardie, 2001). Second, negative citations are often mitigated in academic writing, which blurs the boundary between neutral and polarized citations (MacRoberts & MacRoberts, 1984). We will design new methods to identify mitigations, consulting prior research on automatic identification of negation and modality (Wiegand et al., 2010; Morante & Sporleder, 2012), certainty of scientific claims (Battistelli & Amardeilh 2009; Rubin, 2007), hedges (Medlock & Briscoe, 2007; Szarvas, 2008; Farkas et al. 2010), and speculations (Light et al., 2004).

Associating citation opinions and objects: After identifying the citation opinion purpose, subject, and tones, we need to associate the opinions with their corresponding objects. This is not a trivial task in that the cited work may not be an involved object in a seemingly polarized citation statement. For example, the citation statement *"The x system [1] is important in that …"* contains a positive opinion toward [1], while another seemingly positive statement "*Sentiment classification is important in that …[1]*" does not contain any specific opinion from or toward [1]. The relationship among citation objects can be more complicated when the citation statements span multiple sentences. This problem is further challenging for citations in number format (e.g. "[1]"), which downplays the authors' role by suppressing explicit expressions of relationship between citing and cited work (MacRoberts & MacRoberts, 1984). We will develop new algorithms by studying positional and grammatical relationships between opinions and objects from the training data in order to predict whether an opinion is truly associated with the cited work. The aspect-based opinion mining approaches (e.g. Hu & Liu, 2004; Popescu & Etzioni, 2005) may also help in that they aim for associating customer comments on different aspects of products (e.g. monitor, CPU, and keyboard of a computer).

3.6. Evaluation

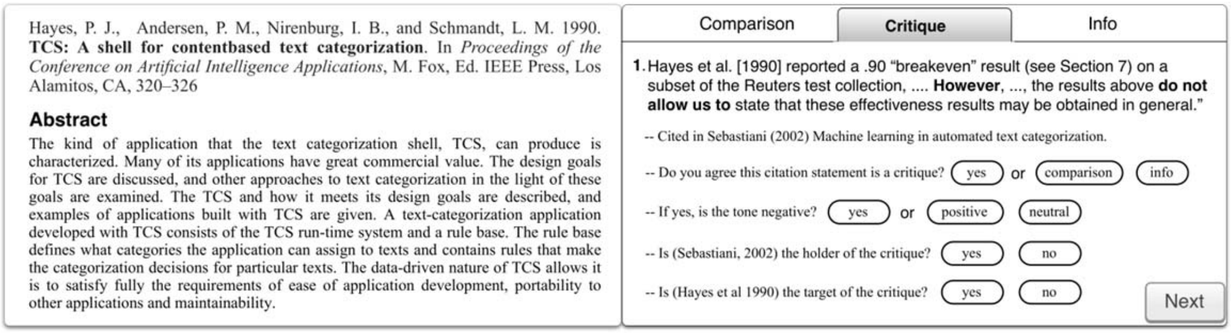
We will conduct three levels of evaluation: (1) algorithm-level evaluation to assess the accuracy of the citation opinion classification models on manually-annotated data; (2) system-level user evaluation in a simulated digital library to test CORA's prediction accuracy and usefulness in assisting researchers with scholarly work; (3) project-wide evaluation to ensure project proceeds rationally.

Algorithm-level evaluation: The focus of this project is to develop accurate predictive algorithms for citation opinion classification. Therefore, our evaluation will focus on algorithm accuracy in the first place. To ensure the reliability of the human-annotated training corpus, we will assign two coders for each article, and measure inter-coder agreement on the annotations (see details of agreement assessment in section 3.4). Predictive models will then be tuned and evaluated on the training corpus using measures like precision and recall in a cross-validation manner.

System-level user evaluation: Since CORA is designed as a plug-in tool for digital libraries, we will create a prototype website with two branches to simulate two real digital libraries (the PubMed website equipped with the PMC-OA corpus; and the ACL Anthology website with the AAN corpus), plug in CORA predictive models, and then recruit biomedical and NLP researchers as users to evaluate: (1) the predictive models' accuracy in classifying the opinions of citations to their works, and (2) the usefulness of CORA in assisting scholarly work.

Considering the large size of the PMC-OA and AAN corpora, we will use a crowdsourcing approach to handle this large-scale evaluation: The email addresses of the cited authors will be extracted from the original articles, and then emails will be sent to invite them to test the simulation website and evaluate the algorithm-predicted results. The crowdsourcing approach is proposed based on the rational assumption that the citees have both the authority of judgment and strong interest in ensuring accurate understanding of citations to their works.

The CORA plug-in works in the following way. For any given article in the PMC-OA or AAN corpora, the system will retrieve the full text of this article and all other articles that cite this article. The citing articles will then be sent to the predictive module to extract and classify citation opinions. All citation opinions will be organized by categories and presented with the full text of the cited article on the prototype website. The citees will be prompted to judge the prediction accuracy, and their responses will be collected for result analysis. See Figure 1 for the interface prototype.



Hayes, P. J., Andersen, P. M., Nirenburg, I. B., and Schmandt, L. M. 1990. **TCS: A shell for contentbased text categorization**. In *Proceedings of the Conference on Artificial Intelligence Applications*, M. Fox, Ed. IEEE Press, Los Alamitos, CA, 320–326

Abstract

The kind of application that the text categorization shell, TCS, can produce is characterized. Many of its applications have great commercial value. The design goals for TCS are discussed, and other approaches to text categorization in the light of these goals are examined. The TCS and how it meets its design goals are described, and examples of applications built with TCS are given. A text-categorization application developed with TCS consists of the TCS run-time system and a rule base. The rule base defines what categories the application can assign to texts and contains rules that make the categorization decisions for particular texts. The data-driven nature of TCS allows it is to satisfy fully the requirements of ease of application development, portability to other applications and maintainability.

1. Hayes et al. [1990] reported a .90 "breakeven" result (see Section 7) on a subset of the Reuters test collection, ... **However**, ..., the results above **do not allow us to** state that these effectiveness results may be obtained in general."

-- Cited in Sebastiani (2002) Machine learning in automated text categorization.

-- Do you agree this citation statement is a critique? (yes) or (comparison) (info)

-- If yes, is the tone negative? (yes) or (positive) (neutral)

-- Is (Sebastiani, 2002) the holder of the critique? (yes) (no)

-- Is (Hayes et al 1990) the target of the critique? (yes) (no)

Figure 1. CORA evaluation interface prototype

In addition to evaluating CORA's prediction performance in simulated digital libraries, we would also like to obtain empirical evidence for CORA's usefulness in solving the aforementioned problems in scholarly communication. Due to copyright issues, some evaluations are not feasible at this time, such as evaluating the completeness of literature review and assessing a biomedical researcher's overall research contribution, because they require analyzing the full-text of all citing articles, many of which may not be open-access. For example, when trying to replicate the citation bias analysis in (Greenberg, 2009), we were not able to replicate the original data set that was collected from PubMed, because more than 80% of the articles were not included in the PMC-OA corpus. Therefore, we limit our user evaluation to the user experience in using our simulation website. We will recruit researchers to use the website for three months and conduct pre-post surveys to examine whether CORA's accuracy is acceptable for researchers, to what extent and in which aspects can CORA assist researchers with their scholarly work. We will also hold an annual workshop to invite local users at Syracuse University and nearby institutions to focus-group discussions on the system performance, website usability design, and potential use scenarios for CORA.

We will particularly compare gender differences in the responses on the usefulness of CORA for assessing research contributions. Our hypothesis is that if citation-count based measures did under-value female STEM faculty's publications, higher interest in qualitative assessment should be observed among female faculty.

Project-wide evaluation: To ensure the project proceeds rationally, the PI has proposed and organized an advisory committee to consult on formative and summative project-wide evaluation on various research and management issues like appropriate staffing, system development, and user study. Two experienced IMLS PIs, Professors Linda Smith (UIUC) and Ruth Small (SU), will mentor the young PI as project execution consultants. Professor Howard Turtle (SU), a senior expert in digital library, information retrieval and natural language processing, will serve as the technical advisor. Professor Anthony Rotolo (SU) will advise as the expert on using the iSchool's social media platform to promote this project, recruit researchers as users, and disseminate research results. Dr. Marie Garland from SU-ADVANCE will serve as liaison for our local user base: the STEM researchers, especially female researchers at Syracuse University and nearby institutions.

3.7. PI's prior research in citation opinion analysis

The PI has conducted two pilot studies on citation opinion analysis. The first one reviewed biomedical researchers' manual practice in citation opinion analysis, which motivated the proposed research design (Yu, 2013b). The second pilot study (Yu and Li, 2013) designed and evaluated a citation opinion classification scheme, which serves as the first version of the framework for annotating citation opinions and conducting automated analysis.

Through past research projects and publications, the PI has demonstrated extensive knowledge of the various research methods required to conduct this research: (1) emotion, sentiment, and opinion analysis in multiple domains, such as customer reviews, political debates, music lyrics, literary novels, and social media (e.g. Zhai et al., 2004; Yu, 2008; Yu et al., 2008a; Hu & Yu, 2011; Yu, 2011; Yu et al., 2011); (2) text classification and feature selection (e.g. Plaisant et al., 2006; Yu, 2008; Shao et al., 2008; Yu et al., 2008b; Yu, 2013a); (3) crowdsourcing for large-scale evaluation (Wang & Yu, 2011; 2012; Yu et al., 2013) ; (4) data annotation and content analysis (Wang & Yu, 2010; Yu & Ku, 2010; Yu et al., 2011)

4. Project resources: personnel, time, budget

Personnel

The PI will participate in every aspect of the project and oversee the team members. A PhD student will be hired to support the PI each year in research, project management, project website maintenance, and writing. Two hourly student assistants will be hired in Year 1 to annotate the training corpus. A research programmer will be responsible for system implementation in Year 2.

Timeline

This project sets the following milestones during the 36-month project period. A detailed timeline is presented in the Schedule of Completion.

- Phase 1 (Month 1-12): recruit PhD student to assist research and project management; recruit and train student assistants to annotate the training corpus.
- Phase 2 (Month 13-18): implement and evaluate a baseline system using knowledge-based methods.
- Phase 3 (Month 18-24): use advanced machine learning and NLP methods to improve the baseline system.
- Phase 4 (Month 25-30): build the simulation website to evaluate the predictive methods and collect user feedback.
- Phase 5 (Month 31-33): analyze user feedback.
- Phase 6 (Month 34-36): project exit, write up final report and system documentation, transfer annotated corpus and source code to a permanent server and open for public downloading for research purpose.

Budget

Requested to carry out this 3-year project is $386,030 An additional $88,979 will be cost shared. The requested funding will (1) support Bei Yu (PI) for project research and development, and dissemination activities; (2) support a doctoral student who will work closely with the PI in research and project management, (3) support hourly graduate assistants to annotate the training corpus in Year 1; (4) support software engineer time in Year 2 to carry out programming tasks; (5) secure server space for three years; (6) support dissemination travel and annual user workshops.

5. Diversity Plan

[Sections 5-7 are particularly enhanced based on reviewers' suggestions from last year.]

As a female faculty member in a STEM field, the PI actively mentors future STEM researchers, especially underrepresented female researchers. This award will give the PI opportunities to further this important goal. The PI has advised or supervised 12 female Master's and doctoral students of diverse ethnic and educational backgrounds in the past four years. The PI has been active in the university-wide Women in Science and

Engineering program (WiSE) and the NSF-funded SU-ADVANCE program, and is a recipient of the SU-ADVANCE cross-sector opportunity travel grant. These programs are dedicated to enhance and support career development for female faculty and students. The PI will continue to use and actively participate in these institutional programs to promote opportunities for underrepresented minority and female students in STEM fields to further their education by engagement with this project as workshop participants or hourly assistants. These programs also hold frequent seminars and workshops to facilitate group discussions among STEM faculty on research and career development. The PI will use these opportunities to introduce CORA to local STEM researchers, recruit them as evaluators, and solicit their feedback on CORA's usefulness in assisting their research and assessing their research impact.

6. Communication Plan

CORA will interact with the research and practice communities in the following ways:

- The research results will be submitted to journals and conference presentations in relevant venues, especially the natural language processing and digital library communities. Target conferences include NLP conferences (ACL, HLT, and EMNLP), library and information science conferences (JCDL, ASIST, and i-Conference), and medical informatics conferences (AMIA and Medicine 2.0).
- The PI will organize a workshop to be affiliated with HLT or JCDL to promote citation opinion analysis as an emerging research area. Additional workshops may be organized depending on the results of the first workshop. The PI has prior experience of co-chairing a workshop on topic-sentiment analysis, and has served on the organization committees of many conferences.
- A publicly accessible website will be constructed to publish news on project progress, and release the annotated corpus and the CORA source code as open-access resources for research purpose. The website will be promoted through the workshop(s), publications, and the iSchool's robust social media platform.
- To increase widespread, the PI will also seek to release the annotated corpus and source code through well-known repositories such as the Linguistic Data Consortium and sourceforge.net.

7. Sustainability Plan

To sustain the project, the PI will use the project website to release the annotated corpus and the CORA source code as open-access resources for the research community, in order to attract more researchers to this research area, and to create an open source community to support continued development of citation opinion analysis software and extension of the training corpus. The website will reside on a permanent server at the PI's home school. The PI will continue to dedicate research time to maintain the project website after the project ends.

To grow the research area of citation opinion analysis, the PI will organize a HLT- or JCDL-affiliated workshop (also described in the communication plan) to invite researchers to submit relevant papers, discussing citation opinion analysis methods and applications. The CORA prototype system will be actively promoted in the biomedical and NLP research communities through collaboration with larger bibliographic databases such as PubMed and Google Scholar.

This project focuses on citation opinion analysis in scientific articles in English. Building on this base, CORA can be extended to more disciplines and languages. The techniques developed can be embedded in software to improve the performance of many applications. More user evaluations will be conducted when sufficient data are available for assessing CORA's ability in assisting literature review, monitoring citation quality, and assessing research contributions, thereby increasing the range of applications in which CORA could be of benefit.

This award would provide invaluable support for the PI as a female faculty member in a STEM discipline. It would help advance the PI's career as an academic researcher and educator, and as a role model for future female students and researchers in library and information science.

Syracuse University
Citation Opinion Retrieval and Analysis (CORA):
An Automated Plug-in Tool for Digital Libraries

Schedule of Completion
Project duration: 3 years

Phase	Period	Deliverables	Month	Activities
1	Year 1	The annotated corpus	0-2	(a) Recruit one PhD student to assist research and project management (b) Recruit two students as annotators (c) Select 500 articles from each of PMC-OA and AAN corpora to be manually annotated; set up GATE annotation environment
			3-5	(a) Train annotators on a small sample. (b) Evaluate inter-coder agreement on a small sample.
			6-7	Organize 1st user workshop to solicit feedback on corpus annotation
			7-12	Annotate all 1,000 articles in the training corpus.
2	Year 2 (first half)	The baseline system	13	develop baseline algorithm for citation statement extraction, evaluate on training corpus
			14-15	Develop baseline algorithm for citation purpose, subject, and tone classification, evaluate on training corpus
			16-17	Develop baseline algorithm to associate citation opinion with the involved objects, evaluate on training corpus
			18	Build simulation website with baseline citation opinion analysis methods plugged in
3	Year 2 (second half)	The further improved system	19-20	Organize 2nd user workshop to demo the baseline system and solicit feedback
			19-20	improve citation statement extraction using advanced techniques like co-reference resolution
			21-22	(a) use the unlabeled data to augment the citation sentiment lexicon (b) use the new lexicon to improve citation opinion classification
			23	Improve citation opinion-object association using advanced techniques like aspect-based opinion
			24	Integrate the improved methods into the simulation website
4	Year 3 (first half)	User feedback	25	Organize 3rd user workshop to demonstrate the improved system and solicit feedback
			26-27	Crowdsourcing large-scale algorithm evaluation by invite all citees in the PMC-OA and AAN corpora to use CORA
			28-30	Begin with inviting a sample of researchers to use CORA for 3 months and conducting pre-survey on system usability; end with post-survey on system usability
5	Year 3 (third quarter)	CORA's evaluation report	31-33	Analyze the user feedback to examine CORA's usefulness in assisting scholarly work
6	Year 3 (fourth quarter)	Final report, system archive, data sharing	34-36	project exit, write up final report and system documentation, transfer annotated corpus and source code to a permanent server and open for public downloading for research purpose.

DIGITAL CONTENT SUPPLEMENTARY INFORMATION FORM

Instructions: This form is required as part of grant applications to the Institute of Museum and Library Services that include activities that create certain types of digital content, such as online collections or databases, metadata, new software tools or electronic systems, or digital research datasets. Your responses to the questions on this form are used by IMLS staff and by expert peer reviewers to better understand technical aspects of your proposed work. Please consult the relevant program guidelines for further instructions on when this form should be included as part of your application.

If you need more space for your response, you may append additional pages as part of the single PDF that you upload with your grant proposal through Grants.gov.

Please indicate which of the following digital products you will create or collect during your project. (Check all that apply):

	If your project will create or collect …	**Then you should complete …**
☒	Born-digital, existing digital, or to-be-digitized content	Part I
☒	New software tools or electronic systems such as databases	Part II
☒	A digital research dataset	Part III

PART I. Projects Creating Digital Content

A. Selection Methodology

A.1 Describe how you will select non-digital materials for digitization.

Not applicable.

A.2 Describe how you will select born-digital or existing digital content for your project collection.

This project will use two corpora, both are academic publications: the PubMed Central Open Access Subset (PMC-OA) includes ~600,000 biomedical articles in .nxml format; the ACL Anthology Network Corpus (AAN) includes ~20,000 open-access articles in natural language processing in .txt format.

B. Converting Non-Digital Materials to Digital Format

B.1 List the types and formats of materials to be digitized and the quantity of each type.

Not applicable.

B.2 List the equipment and software that you will use to digitize each of these formats or the name of the digitization services provider who will perform the work.

Not applicable.

B.3 List the digital file formats (e.g., TIFF, JPEG, MPEG) that you will produce during the digitization work and the anticipated quality standards for each file format (e.g., resolution, bit-depth, color/grayscale, pixel dimensions, sampling rate).

Not applicable.

B.4 If different digital versions of content will be created during the digitization process (e.g., preservation master, access copy, thumbnail) list the type, format, and number of each version.

Not applicable.

C. *Repurposing Existing Digital Content or Creating New Digital Content*

C.1 List the types and formats of born-digital or existing digital content that you will create or repurpose and the quantity of each.

New markups with regard to citation purpose, subject, tone, and objects will be added to the original articles as inserted tags. The annotated articles will be stored as xml files. All citation statements in the original corpora will be marked up (~620,000 articles).

C.2 If you will be creating new born-digital content or converting existing digital content to new formats, list the equipment and software that you will use to create each of these formats or the name of the services provider who will perform the work.

Human annotators will use GATE to annotate articles and export the annotations to xml files. GATE is an open-source Natural Language Processing toolkits. It provides a graphical user interface for annotators to annotate the content of text documents and export the annotations in XML format.

The research team will then write computer programs (Perl, Python, Java, etc.) and use open-source NLP and machine learning packages (NLTK, OpenNLP, SVM-Light, Weka, etc.) to learn patterns from human annotations and use the patterns to automatically mark up the original articles.

C.3 If you will be converting existing digital content to new formats, list the new digital file formats and relevant information on the anticipated quality standards (e.g., sampling rate, pixel dimensions).

New markups with regard to citation purpose, subject, tone, and objects will be added to the original articles as inserted tags. The annotated articles will be stored as xml files. All citation statements in the original corpora will be marked up (~620,000 articles). Quality of human annotations will be measured by inter-coder agreement; quality of computer annotations will be measured by comparing against human annotations.

C.4 If different versions of digital content will be created during the conversion or re-purposing process (e.g., preservation master, access copy, thumbnail), list the type, format, and number of each different version.

not applicable

D. Digital Workflow and Asset Maintenance/Preservation

D.1 Describe your quality control plan.

Quality of human annotations will be measured by inter-coder agreement; quality of computer annotations will be measured by comparing against human annotations.

D.2 Describe your plan for preserving and maintaining digital assets during and after the grant period (e.g., storage systems, data standards, technical documentation, migration planning, commitment of organizational funding for these purposes).

The CORA source code and the annotated corpus (consisting of XML files) will be moved to a permanent server maintained by the IT Support department at the School of Information Studies. The maintenance cost of this server is covered by the School of Information Studies.

E. Metadata

E.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

A self-developed annotation scheme will be used to annotate citation opinion. See details in the research design section in the Narrative. The annotations will be inserted into the original articles as extra xml tags.

E.2 Describe how you will use metadata to enhance the management, discovery, and use of your digital content.

The annotation scheme serves as the foundation for the citation opinion analysis by defining the citation categories that the automated tool is expected to classify automatically.

E.3 Explain your strategy for preserving and maintaining metadata created and/or collected during your project, during and after the grant period.

During the project, the annotations and computer codes will be stored on the virtual server budgeted for this project. The School's ITS conducts regular backup for the entire server. All data is stored on disk, all disks are backed up nightly (restore point). Local backups (restore points) are retained for 30 days. A weekly backup set is sent to off-site storage for disaster recovery purposes only. Off-site backups are retained for 30 days.

E.4 Explain what metadata-sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content created or repurposed during your project (e.g., an Advanced Programming Interface or other support to allow batch queries and retrieval of metadata).

The annotations and CORA source code will be open-access for research purpose. It will reside on a server maintained in the School of Information Studies at Syracuse University. It will be downloadable through web browsers. Collaborations with the Qualitative Data Repository (QDR) and the Linguistic Data Consortium (LDC) will be sought to deposit an electronic copy of the annotated corpus to LDC for public access.

F. Copyright and Intellectual Property Rights

F.1 Explain the current copyright or intellectual property status of the content you intend to digitize, create, or repurpose. Describe the quantity or percentage of materials that are in the public domain and/or have restrictions that will require you to obtain permissions. If you have already obtained permission to use and provide public access to materials under copyright or other restrictions, provide the quantity of such materials, and the documentation you possess granting such permissions.

The content that will be repurposed is currently 100% in the public domain, free for research use under creative commons agreement. The research team owns the copyright for the new annotations and computer programs. The entire annotated corpus and computer source code will be open-access for research purpose under the Creative Commons agreement. The annotated corpus and computer source code will reside on a web server for free download.

F.2 If you will need to obtain permissions during your project, describe the process you will use to request and obtain them.

Not applicable.

F.3 Are there any materials you will be digitizing, creating, or repurposing that may raise privacy concerns? If so, what is your plan for addressing them?

No.

F.4 If your project will include online users or others outside your organization contributing metadata, social media comments, or other content to your digital resources, describe your plan to obtain releases or permissions from these content contributors. What rights and permissions will you require such contributors to transfer to your organization?

We will ask online users (citees) to judge the accuracy of the computer-predicted citation opinion categories. We will post a user agreement form when a user enters our website. Users can choose two levels of participation: (1) provide the judgments but don't release them to the public; (2) provide the judgments and allow them to be added to the annotated corpus.

We will also send surveys to researchers who use our website with regard to the usefulness of the website. Permission form for data collection will be displayed on the screen before users take the surveys. Survey results will be published in aggregated form only.

For both user studies , we will apply for IRB approval in Syracuse University.

G. Access And Use

G.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

The original corpora are open-access for research purpose. The repurposed material with added manual- or computer-annotations for citation statements, will also be open-access for research purpose. We will create a website to make the repurposed material and new markups available for public downloading via the Internet.

G.2 We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in openly accessible journals, depositing works in openly accessible repositories, and using non-restrictive licenses such as the "CC Zero – No Rights Reserved" that dedicate digital content to the public domain. What ownership rights will your organization assert over the new digital content, and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users of the digital resources.

The research team owns the copyright for the new annotations and source code. The entire annotated corpus and source code will be open-access for research purpose by following the Creative Commons agreement. The annotated corpus and source code will reside on a web server for free download.

G.3 Provide URL(s) for any examples of previous digital collections or content your organization has created.

repurposed Senatorial speech from the 101st to 109th Congress, downloaded from Thomas.gov, speeches extracted reorganized by speakers and dates, available for public downloading : http://textmining.syr.edu/beiyu/Senate-compressed/

Part II. Projects Creating Software Tools and Electronic Systems

A. General Information

A.1 Describe the software tool or electronic system you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) the system or tool will serve.

This project will create Citation Opinion Retrieval and Analysis [CORA], an automated tool that can be plugged into a full-text bibliographical database, retrieve the context of each citation and categorize them by citation purpose, subject, and polarity. CORA, will save librarians and researchers significant amount of time to find the most useful comments from a large number of citations. CORA will also provide a new, qualitative approach for assessing research impact. CORA can also help monitoring the quality of scientific publications by facilitating easier identification of citation bias and inaccurate citations from the reorganized citations.

A.2 List other existing digital tools that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

CORA is an innovative tool - no existing digital tools wholly or partially perform the same function.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your new digital content.

The research team will write computer programs (Perl, Python, Java, etc.) and use open-source NLP and machine learning packages (NLTK, OpenNLP, SVM-Light, Weka, etc.) to automatically mark up the citation opinions in the original corpora.

B.2 Describe how the intended software or system will extend or interoperate with other existing software applications or systems.

CORA is designed as a plug-in tool that can be integrated into any bibliographical databases, enhancing digital libraries with citation opinion analysis functions. CORA uses XML as input and output format, which facilitates convenient data management and transfer, and optimal system interoperability.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software or system you will create.

Windows and Linux file systems

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software or system.

The CVS version control system will be installed on the project server to facilitate version control of source code during system development. A user manual will be prepared and released together with the final version of source code.

B.5 Provide URL(s) for examples of any previous software tools or systems your organization has created.

C. <u>Access and Use</u>

C.1 We expect applicants seeking federal funds for software or system development to develop and release at least a beta version of these products as open-source software. What ownership rights will your organization assert over the new software or system, and what conditions will you impose on the access and use of this product? Explain any terms of access and conditions of use, why these terms or conditions are justifiable, and how you will notify potential users of the software or system.

The CORA source code will be open-access for research purpose.

C.2 Describe how you will make the software or system available to the public and/or its intended users.

We will create a project website to release the source code after project ends.

Part III. Projects Creating Digital Research Data (Data Management Planning)

We expect exemplary management and sharing of research data. The purpose of this part of the form is to help us understand your research practices and plans for management of data that will be generated through your project. Please address each question that applies to your proposed project.

1. Summarize the intended purpose of the research, the type of data to be collected or generated, the approximate dates when the data will be generated or collected, and the anticipated volume of data.

This research aims to develop CORA, an automated tool for citation opinion retrieval and analysis. It will annotate the citation statement purpose, subject, polarity, and objects of 620,000 full-text articles in biomedical and NLP disciplines. A training corpus of 1,000 articles will be manually annotated in the first year of the project. The remaining articles will be annotated by computer programs. Each annotated article is ~100KB; the entire corpus will be ~62GB

2. Does the proposed research activity generating the dataset(s) require approval by any internal or institutional review panel? If so, has the proposed research activity already been approved? If not, what is your plan for securing approval?

We will ask online users (citees) to judge the accuracy of the computer-predicted citation opinion categories. We will post a user agreement form when a user enters our website. Users can choose two levels of participation: (1) provide the judgments but don't release them to the public; (2) provide the judgments and allow them to be added to the annotated corpus.

We will also send surveys to researchers who use our website with regard to the usefulness of the website. Permission form for data collection will be displayed on the screen before users take the surveys. Survey results will be published in aggregated form only.

For both user studies , we will apply for IRB approval in Syracuse University, and to the best of our knowledge, we believe these user studies belong to the exempt category.

3. Will you collect any confidential or private information about individuals (e.g., names, contact information, health status) or proprietary information about organizations? If so, detail the specific steps you will take to protect such information while you prepare the research data files for public release.

No, we will not collect any confidential or private information. We will contact researchers as potential system users by automatically extracting their email addresses from their open-access publications, which are counted as public information.

4. If you will collect additional documentation such as consent agreements or signed certifications along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

Consent forms will be displayed in the web browser before users enter their responses. The consent form will be included as part of the annotated corpus for open access.

5. How will you manage intellectual property interests related to the dataset(s)? Who will claim ownership of the intellectual property rights related to the dataset(s)? How will those claims of ownership be communicated to others?

The research team owns the copyright to the annotations and computer source code. They will be released for free for research purpose under Creative Commons agreement.

6. Which technologies, instruments, or tools will you use to collect or generate the data? Provide details about hardware or software; electronic formats for data capture or storage; standards or local practices for data content and encoding; controlled vocabularies or other mechanisms for data normalization and consistency; and any other relevant technical requirements or dependencies for understanding, retrieving, displaying, or processing the dataset(s). If the data will be encrypted at any point in its active or inactive life, explain the reasons for choosing to encrypt the data and how the decryption key will be stored, protected, and made available if necessary.

Human annotators will use GATE annotation tool to generate annotations. GATE is an open-source Natural Language Processing toolkits. It provides a graphical user interface for annotators to annotate the content of text documents and export the annotations in XML format. Research programmers will write computer algorithms that learn patterns from human annotations to automatically generate annotations. The algorithms will be developed based on open-source machine learning and NLP software, such as Weka, SVM-Light, OpenNLP, NLTK, etc.

7. What metadata will you capture or create along with the dataset(s)? What standards or schema will you use to express the metadata? Where will the metadata be stored, and in what format(s)? How will you permanently associate and manage the metadata with the dataset(s) it describes?

A self-developed annotation schema will be used to annotate citation purpose, subject, tone, and objects. See details of the annotation schema in the research design section in the Narrative. The annotations will be inserted into the original articles as extra xml tags. The annotated corpus will be open-access and downloadable from the project website.

8. During the research project, where will the data and metadata be stored and on what type of media? Who will have access to the data and/or copies of the data during the project? How many backup copies will you maintain during the project, and how frequently will you refresh the backup copies? Who will be responsible for data backup? Where will you store the backup copies of the data and metadata during the project?

During the project, the annotations and computer code will be stored on the virtual server budgeted for this project. The School's ITS conducts regular backup for the entire server. All data is stored on disk, all disks are backed up nightly (restore point). Local backups (restore points) are retained for 30 days. A weekly backup set is sent to off-site storage for disaster recovery purposes only. Off-site backups are retained for 30 days.

9. Once the research project is completed, what is the long-term plan for archiving, managing, and making the metadata and dataset(s) available? What steps will you take to prepare the data for sharing (e.g., labeling missing data, standardizing measures statistical disclosure limitation methods)?

The CORA source code and the annotated corpus (consisting of XML files) will be moved to a permanent server maintained by the IT Support department at the School of Information Studies. The maintenance cost of this server is covered by the School of Information Studies. The data set will be open-access for research purpose. A user manual will be created and released to explain the meaning of the annotations and source code.

10. Identify where you will be depositing research dataset(s) and metadata into:

a) an institutional repository:

Name:__________ URL: __________

b) a subject specific research community digital repository:

Name:__________ URL: __________

c) or some other publicly accessible repository:

Name: the PI's server at home school URL: http://beiyu.syr.edu

Does this repository enforce any access restrictions? □ Yes (If yes, describe.) or ☒ No

If so, how will they be mitigated to allow the public free access to these data? Detail the experience this repository has in managing research datasets and metadata with similar attributes? What preservation and backup procedures does this repository use?

The PI's web server is expected to provide free access to both annotated corpus and source code by facilitating free http or ftp downloading from the web server located at the PI's home school. The PI has used this website to release other corpus previously.

To increase visability, the PI has also considered other digital repositories like
(1) the Qualitative Data Repository at SU (QDR) https://www.maxwell.syr.edu/news.aspx?id=77309421232
(2) the Linguistic Data Consortium at UPenn (LDC) http://www.ldc.upenn.edu/
to host the annotated corpus. QDR is still under construction and LDC requires membership for free access. Negotiations will be sought to waive the non-membership access charge for our corpus. Online hosts for open-source software such as sourforge.net will be a candidate for depositing the computer code for open access

11. When and how frequently will you review this data management plan? How will the implementation be monitored?

The data management plan will be reviewed at the following check points:
- the training corpus of 1,000 articles are annotated
- the baseline system is constructed
- the improved system is constructed
- user feedback is collected
- project ends

We will release available data, annotated corpus or computer code, once they are ready for public use, by posting them to the above repositories.