CEUR-WS.org/Vol-2253/paper11.pdf

Identifying Citation Contexts: a Review of Strategies and Goals.

Agata Rotondi, Angelo Di Iorio, Freddy Limpens
Department of Computer
Science and Engineering
University of Bologna, Italy
agata.rotondi@unibo.it
angelo.diiorio@unibo.it
freddy.limpens@unibo.it

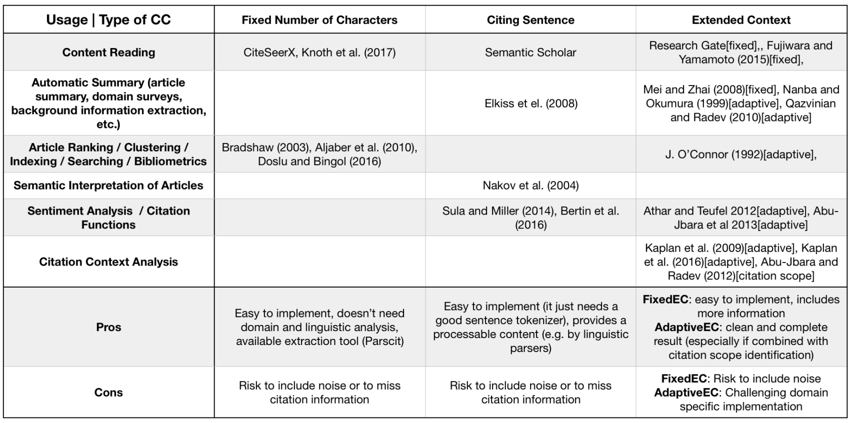
Abstract

English. The Citation Contexts of a cited entity can be seen as little tesserae that, fit together, can be exploited to follow the opinion of the scientific community towards that entity as well as to summarize its most important contents. This mosaic is an excellent resource of information also for identifying topic specific synonyms, indexing terms and citers' motivations, i.e. the reasons why authors cite other works. Is a paper cited for comparison, as a source of data or just for additional info? What is the polarity of a citation? Different reasons for citing reveal also different weights of the citations and different impacts of the cited authors that go beyond the mere citation count metrics. Identifying the appropriate Citation Context is the first step toward a multitude of possible analysis and researches. So far, Citation Context have been defined in several ways in literature, related to different purposes, domains and applications. In this paper we present different dimensions of Citation Context investigated by researchers through the years in order to provide an introductory review of the topic to anyone approaching this subject.

Italiano. *Possiamo pensare ai Contesti Citazionali come tante tessere che, unite, possono essere sfruttate per seguire l'opinione della comunità scientifica riguardo ad un determinato lavoro o per riassumerne i contenuti più importanti. Questo mosaico di informazioni può essere utilizzato per identificare sinonimi specifici e Index Terms nonchè per individuare i motivi degli autori dietro le citazioni. Identificare il Contesto Citazionale ottimale è il primo passo per numerose analisi e ricerche. Il Contesto Citazionale è stato definito in diversi modi in letteratura, in relazione a differenti scopi, domini e applicazioni. In questo paper presentiamo le principali dimensioni testuali di Contesto Citazionale investigate dai ricercatori nel corso degli anni.*

1 Introduction and Background

Researchers consider as Citation Context (CC) different snippets of text around a citation marker. These differences of width influence the applications that exploit CC as source of information. For example, Qazvinian and Radev (2010) showed that using also implicit citations (i.e. sentences that contain information about a specific secondary source but do not explicitly cite it) for generating surveys, rather than citing sentences alone, improve the results. Ritchie et al. (2008) compared different widths of CC in order to find the most appropriate window for identifying Index Terms. They proved that varying the context from which the Index Terms are gathered has a significant effect on retrieval effectiveness. Aljaber et al. (2010) tested different sizes of CC for a document clustering experiment. They claimed that a window size of 50 words from either side of the citation marker works better than taking 10 or 30 terms or the citing sentence alone, whatever its size is. From their analysis, relevant synonymous and related vocabulary extracted from this window of text, in combination with an original full-text representation of the cited document, are effective for document clustering. We can claim that the issue of finding the optimal CC for a specific application is a challenging task that interests researchers and which is at the base of every study that exploits the CC as a source of information.



Usage | Type of CC	Fixed Number of Characters	Citing Sentence	Extended Context
Content Reading	CiteSeerX, Knoth et al. (2017)	Semantic Scholar	Research Gate[fixed],, Fujiwara and Yamamoto (2015)[fixed],
Automatic Summary (article summary, domain surveys, background information extraction, etc.)		Elkiss et el. (2008)	Mei and Zhai (2008)[fixed], Nanba and Okumura (1999)[adaptive], Qazvinian and Radev (2010)[adaptive]
Article Ranking / Clustering / Indexing / Searching / Bibliometrics	Bradshaw (2003), Aljaber et al. (2010), Doslu and Bingol (2016)		J. O'Connor (1992)[adaptive],
Semantic Interpretation of Articles		Nakov et al. (2004)	
Sentiment Analysis / Citation Functions		Sula and Miller (2014), Bertin et al. (2016)	Athar and Teufel 2012[adaptive], Abu-Jbara et al 2013[adaptive]
Citation Context Analysis			Kaplan et al. (2009)[adaptive], Kaplan et al. (2016)[adaptive], Abu-Jbara and Radev (2012)[citation scope]
Pros	Easy to implement, doesn't need domain and linguistic analysis, available extraction tool (Parscit)	Easy to implement (it just needs a good sentence tokenizer), provides a processable content (e.g. by linguistic parsers)	**FixedEC**: easy to implement, includes more information **AdaptiveEC**: clean and complete result (especially if combined with citation scope identification)
Cons	Risk to include noise or to miss citation information	Risk to include noise or to miss citation information	**FixedEC**: Risk to include noise **AdaptiveEC**: Challenging domain specific implementation

Figure 1: Survey Summary

1 With the purpose of providing a useful background to anyone approaching this question, in the following sections we give an overview of different dimensions of textual CC investigated in literature. We classified them in 3 main categories: a) fixed number of characters b) citing sentence c) extended context (fixed and adaptive), and we summarized our analysis in Figure1. We focus on the strategies to identify the correct textual CC of a citation, nevertheless other CC related topics have been investigated in literature as for example citation recommendations (see Farber (2018) and Ebesu (2017))

The belief of the need of a clear introductory survey about how CC has been differently shaped in literature came to our mind when we faced the problem of defining the optimal CC for the Semantic Coloring of Academic References (SCAR) project[1] (Di Iorio et al., 2018). The goal of the SCAR project is to enrich bibliographies of scientific articles by adding explicit meta data about individual bibliographic entries and to characterize these entries according to multiple criteria. With this purpose, we are studying a set of properties to support the automatic characterization of bibliographic entries and one of our primary source of information is the textual content around citation markers, i.e. the CC. We are currently investigating on finding the best span of text for our needs. By reviewing the literature, we realized that different approaches correspond to different tasks and are also related to the linguistic domain of application. The SCAR project as well as this review are focused on the English language but it would be interesting to extend this study to other languages.

2 Fixed Number of Characters

A good way to start exploring how the CC can be diversely defined is to look for well known examples. One of these is the public search engine and digital library for scientific and academic papers CiteSeerX[2]. This web platform allows users to browse papers' references and to read the context in which a reference is cited. The function enables the reading of 200 characters before and after the citation marker. Here the choice of the CC width is not directly related to further analysis and applications as the purpose is the mere reading of text by users. As Ii et al. (2014) describe, CiteSeerX uses ParsCit (Councill et al., 2008) for citation extraction. ParsCit is a freely available, open-source implementation of a reference string parsing package which performs reference string segmentation and CC extraction. The size of the context is configurable, but by default extends to 200 characters on either side of the match. ParsCit is a well know software and is used in different projects. For example, the Association Of Computational Linguistics (ACL) Anthology Network[3] uses ParsCit for curation. Doslu and Bingol (2016) also used ParsCit in their work regarding how to rank articles for a given topic. The authors exploited the information contained in the CC of a certain paper for detecting important articles and providing focused directions to access the literature about a topic. They stated that the words that are used to describe a cited paper stand close to the citation marker, and this is their motivation for choosing a fixed window size context. Before Doslu and Bingol, also Bradshaw (2003) used CC to index cited

[1] http://dasplab.cs.unibo.it/index.php/scar/

[2] http://citeseerx.ist.psu.edu/index

[3] http://aan.how/index.php/home/about

paper for specific topics. He designed the Reference Direct Indexing in which measures of relevance and impact are joined in a single retrieval metric based on the comparison of the terms authors use in multiple CC of a document. The CC Bradshaw used to index the documents are directly gathered from CiteSeerX. Also the tool presented by Knoth et al. (2017), who address the problem of automatically retrieving and collecting CC for a given unstructured research paper, extract a CC window of fixed length corresponding to 300 characters before and after a citation marker. The approach of considering as CC a fixed length snippet around the citation marker is a naive baseline method. It can be used to retrieve terms related to a cited entity and the accuracy of applications that employ it might be improved for example by considering sentence or paragraph boundaries(Aljaber et al., 2010). This kind of context is unsuitable if the CC needs to be further analyzed, for example by using syntactic parsers, or if its content have to be represented in a coherent formal way where the meaning and structure of sentences have to be preserved.

3 Citing Sentence

Another famous platform among scholars is Semantic Scholar[4]. This subjective search service for journal articles provides several functions for browsing papers among which the possibility of quickly read the CC of each citation. This service allows reading more than one excerpt of text for each entity (when available). Each CC shown corresponds exactly to a citing sentence, i.e. the sentence that contains the targeted reference marker. Implicit citations[5] are also investigated by exploiting lexical hooks and also in these cases the CC excerpts shown are in the form of a full sentence. The same CC window has been adopted in several projects. Nakov et al. (2004) investigated the use of CC for semantic interpretation of bioscience articles. Starting from the collection of the citing sentences related to a specific cited entity (that they call *citances*), they used the output of a dependency parser to build paraphrase expressing relations between two named entities. As commented before, parsers need to be fed with full sentences in order to provide proper representations and this work is a clear example where a fixed length CC would not have been an appropriate input. Also Elkiss et al. (2008) focused their research on the set of citing sentences of a given article (named by the authors *citation summaries*) testing the biomedical domain. Despite Elkiss study did not rely on any strictly sentence based technique (they employed cosine similarity and tf-idf), both their hypothesis are grounded on the importance of citing sentences boundaries. Sula and Miller (2014) presented an experimental tool for extracting and classifying citation contexts in humanities. Their approach is based on citing sentences from which they extracted features (e.g. location in document) and polarity (evaluating n-grams with a naive Bayes classifier). Bertin et al. (2016) followed a similar approach to identify n-grams and sentiment in CC. They chose to work on a sentence basis stating that sentences are the natural building blocks of text and likely to include the context of a specific reference. Starting from citing sentences they extracted 3-grams containing verbs, together with position in the paper and type of section according to the IMRaD structure in order to analyze the combination and distribution of these features in the biomedical domain. Citing sentence as a base unit for CC is mostly chosen in hard sciences domains. In fact, scientific communities have particular ways of using language and specific conventions that reveal clear disciplinary differences. Hyland (2009) describes some of these language variations that go from terminology differences to different citations practices and rhetorical preferences. Writers use different sets of reporting verbs to refer to others work (engineers *show*, philosophers *argue*, biologists *find* and linguists *suggest*); frequencies of hedges and self citations, directives and n-grams also diverge across fields. In the humanities writers tend to include extensive referencing and build a background for the heterogeneous readership while in hard sciences most of the readers share a common context with writers. This attitude clarifies citers' behaviors in different domains and makes us presume that CC in humanities might be more complex than in hard sciences. Following these considerations, it is reasonable to con-

[4] https://www.semanticscholar.org

[5] More in details, with implicit citations we refer to those mentions of a work where the relation cited entity-citing entity is not provided by a citation marker but rather by a lexical object related to the cited entity. E.g.: *The heuristics based on **WordNet** and Wikipedia ontologies are very sensitive to preprocessing* is an implicit citation of George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

clude that for choosing the appropriate CC width one needs to take into account not only the task he is going to face but also the domain of applications and the specificity of the language. In this sense, CC as citing sentence might not always correspond to the entire fragment of text referring to a targeted citation marker.

4 Extended Context

Extending CC beyond the citing sentence can prove useful in many cases as illustrated by the social networking site for researchers ResearchGate[6]. Every document in this platform's database can be inspected according to different prospectives. Among them, readers can browse documents citations lists and access CC (when available) displayed in the form of: 1 sentence before the citing sentence + citing sentence + 1 sentence after the citing sentence. This window size allows users to better understand the full context of a citation without loosing any possible informations contained in the nearby sentences. This is particularly relevant for the task of polarity identification of citations. Athar and Teufel (2012) have shown that authors' sentiments are most likely expressed outside the citing sentences. Sentiment in citations is often hidden and especially criticism might be hedged both for politeness and for political reasons (MacRoberts and MacRoberts, 1984). Citing sentences are typically neutral and in particular negative polarity occurs in the following sentences (Teufel et al., 2006), see for example (from (Platt, 1990)):

In [19, sec. 11.11], Vapnik suggests a method for mapping the output of SVM to probabilities by decomposing the feature space []. Preliminary results for this method, are ***promising****.However, there are some* ***limitations*** *that are overcome by the method of this chapter.*

Particularly for, but not limited to, polarity identification tasks, a context extended to the nearby sentences can supply the complete set of information about a citation to applications and readers. Sentences nearby a citing sentence can be add as part of the CC according to a fixed schema or by following an adaptive approach.

4.1 Fixed Extended Context

Besides ResearchGate and the aforementioned Ritchie's work, who studied different window sizes of CC for identifying Index Terms, also Mei and Zhai (2008) implemented a fixed extended context for their study of summarizing articles influence. For their impact-based summarization task they used a 5 sentences window size, with 2 sentences before and after the citing sentence. This technique allows to include more info in the CC but at the same time the risk of adding noise is high. This is why most of the literature concerning extended CC rather provides adaptive methods.

A mention is needed to the work of Fujiwara and Yamamoto (2015), mostly for their overall project than for the CC retrieval approach which relies on a very basic technique (they include the sentence after the citing one if the reference marker is at the end of the citing sentence and limit long citing sentences to 240 characters before and after citation markers). The authors built the Colil database where CC of the life sciences domain are stored, and made it available to users through a web-based search service. For each resource stored in the database, a list of CC in which the resource has been cited is returned to the user who can easily read how a work is perceived and used by different authors.

4.2 Adaptive Extended Context

O'Connor (1982) was the first who investigated the CC as a sequence of sentences - a multi-sentence citing statement. His purpose was to study the words of CC as possible improvement for the retrieval of the related cited entities. He wrote 16 complex and detailed computer rules (not completely computer procedures at that time) with linguistic, structural and more general features for the selection of citing statements. Nanba and Okumura (1999) presented a system to support writing surveys of a specific domain. They see the CC as a succession of sentences where the possible connections are indicated by 6 kinds of cue words (anaphora, negative expression, 1st and 3rd person pronoun, adverb, other) that they use for retrieving the suitable CC for their system. To identify the full span of CC, Kaplan et al. (2009) presented a different method based on co-reference chains. They built a SVM (Cortes and Vapnik, 1995) classifier with 13 features (among which: cosine similarity, gender and number agreement,

[6] https://www.researchgate.net

semantic class agreement etc.) that are tested in order to find the best configuration. Results of the classifier alone and in combination with cue-based techniques are promising. Despite the little data analyzed for the project, Kaplan raised some interesting remarks about CC. Particularly, they stated that sentences of CC are not necessarily contiguous. Qazvinian and Radev (2010) explored the task of retrieving background information close to explicit citations by implementing a probabilistic inference model (Markov Random Field). Like previous authors, they observed that the majority of sentences related to a citation directly occur after or before the citation or another context sentence; however they also confirmed Kaplan's intuition about possible gaps between sentences describing a cited paper. Athar and Teufel (2012) tried to go further by attempting to retrieve all the mentions of a cited entity within the full text of the citing paper. As claimed by the authors, mentions to a cited entity can occur in the full article and are necessary to identify the real sentiment toward the cited work. Their first experiment of manual annotation proved the insight that retrieving all the mentions of a cited entity increases citation sentiment coverage. Also the SVM framework implemented by the authors, despite limited to a 4 sentence window, outperformed a single sentence baseline system. Abu-Jbara et al. (2013), with the purpose of adding qualitative aspects to standard quantitative bibliometrics (H-Index, G-Index, etc.), analyzed the text surrounding a citation in order to define the citer's purposes and polarity. This piece of text (CC), is retrieved with a sequence labeling method. Starting from the citing sentence, Abu-Jbara's team used CRF (Lafferty et al., 2001) to determine if the sentence before and the two sentences after the citing sentences have to be included in the CC. The features for the CRF model are both structural (e.g. position of the current sentence with respect to the citing sentence) and lexical (e.g. presence of demonstrative determiners). Kaplan et al. (2016) named Citation Block Determination(CBD) the task of detecting non-explicit citing sentences and faced it by testing various features representing different aspects of textual coherence. Non local mentions are excluded from what they formalized as a binary classification task of sentences from the citing one. They tested different relational and entity coherence features and their combinations. Experiments showed that the CRF method fits better the task than the SVM approach.

The different works briefly described so far give an overview of the most interesting techniques explored by researchers. From rule-based approaches to probability methods, the implemented features are most of the time domain-specific relying on particular vocabulary and on stylistic and rhetorical habits.

4.2.1 Citation Scope

Related to the Adaptive Extended Context topic is the identification of the Scope of a citation. So far we have discussed different ways of including in the CC what is outside the citing sentence but at the same time related to it. The idea is to extend the context. However, there are cases in which the citing sentence does not completely refer to the targeted citation or where the context of multiple citations overlap. In these cases the aforementioned approaches of CC extraction would include noise and affect applications results. See for instance the following example where the whole citing sentence might produce a negative polarity despite the neutral value of the citation:

The negative results produced by the BoW approach led our team to change direction and we tested a SVM(CORTES, 1995) classifier.

Finding a procedure to cut out the precise scope of a citation is a tricky and challenging task for which little experiments have been done.

Athar (2011) suggested to trim the parse tree of each citing sentence and to keep only the deepest clause in the subtree of which the citation is a part. Abu-Jbara and Radev (2012) explored 3 different methods for identifying the scope: word classification, sequence labeling and segment classification. Results showed that the scope of a given reference consists of units of higher granularity than words. In fact, the segment classification technique achieved the best performance. Despite the interesting results, we agree with Hernandez and Gomez (2016) who stated that additional work is required to improve the citation scope identification task. The need of further research in this field is also encouraged by the analysis of Jha et al. (2017) who performed an annotation experiment on a sample of the ACL Anthology Network revealing that, on average, the reference scope for a given target reference contains only 57.63 per cent of the original citing sentence.

5 Conclusion

We have reviewed what we consider the most interesting works about CC identification in order to provide a solid background to anyone interested in the topic and especially to those researchers who are facing the task of identifying the best approach for their studies. We did not compare the different strategies with the purpose of ranking them, but we rather showed that there exists various relations between a methodology and the usage, domain, and language specificity of its possible applications.

References

Abu-Jbara, A., and Radev, D. 2012. *Reference Scope Identification in Citing Sentences*. In Proc. of NAACL HLT, (p. 80-90).

Abu-Jbara, A., Ezra, J., and Radev, D. 2013. *Purpose and Polarity of Citation: Towards NLP-based Bibliometrics*. In Proc. of NAACL HLT, (p. 596-606).

Aljaber, B., Stokes, N., Bailey, J., and Pei, J. 2010. *Document Clustering of Scientific Texts Using Citation Contexts*. Information Retrieval, 13, (p.101-131).

Athar, A. 2011. *Sentiment Analysis of Citations using Sentence Structure-Based Features*. Proceedings of NAACL-HLT, (p.81-87).

Athar, A., and Teufel, S. 2012. *Detection of Implicit Citations for Sentiment Detection.*. In Proc. of DSSD, (p. 18-26).

Bertin, M., Atanassova, I., Sugimoto, C., and Lariviere, V. 2016. *The Linguistic Patterns and Rhetorical Structure of Citation Context: an Approach Using N-Grams*. Scientometrics, 109(3).

Bradshaw, S. 2003. *Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes*. In Proc. of ECDL, (p. 499-510).

Cortes, C. and Vapnik V. 1995. *Support-Vector Networks*. Machine Learning, 20 (3), (p. 273-297).

Councill, I., Giles, C., and Kan, M. 2008. *ParsCit : An Open-Source CRF Reference String Parsing Package*. In Proc. of LREC, (p. 661-667).

Di Iorio, A., Limpens, F., Peroni, S., Rotondi, A., and Tsatsaronis, G. 2018. *Investigating Facets to Characterise Citations for Scholars*. In Proc. of SAVE-SD Workshop.

Doslu, M., and Bingol, H. 2016. *Context Sensitive Article Ranking with Citation Context Analysis*. Scientometrics, 108 (2), (p. 653671).

Ebesu, T., and Fang, Y. 2017. *Neural Citation Network for Context-Aware Citation Recommendation*. In Proc. of SIGIR, (p. 10931096).

Elkiss A., Shen S., Fader A., Erkan G., States D., and Radev D. 2008. *Blind Men and Elephants: What Do Citation Summaries Tell Us About a Research Article?*. American Society for Information Science and Technology, 59 (1), (p. 51-62).

Farber M., Thiemann A., and Jatowt A. 2018. *To Cite, or Not to Cite? Detecting Citation Contexts in Text*. In Proc. of ECIR: Advances in Information Retrieval, (p. 598-603).

Fujiwara, T., and Yamamoto, Y. 2015. *Colil: a Database and Search Service for Citation Contexts in the Life Sciences Domain*. Biomedical Semantics, 6(38).

Hernandez-Alvarez, M., and Gomez, J. 2016. *Survey About Citation Context Analysis: Tasks, Techniques, and Resources*. Natural Language Engineering, 22(3), (p. 327-349).

Hyland, K. 2009. *Writing in the Disciplines: Research Evidence for Specificity*. Taiwan International ESP Journal, 1(1), (p. 5-22).

Jha, R., Jbara, A., Qazvinian, V., and Radev D. 2017. *NLP-driven Citation Analysis for Scientometrics*. Natural Language Engineering, 23(1), (p. 93-130).

Lafferty, J., McCallum, A., and C.N. Pereira, F. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In Proc. of ICML, (p. 282-289).

Ii, A., Wu, J., and Giles, C. 2014. *CiteSeerX : Intelligent Information Extraction and Knowledge Creation from Web-Based Data*. In Proc. of AKBC, (p. 1-7).

Kaplan, D., Iida, R., and Tokunaga, T. 2009. *Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-Chain Based Approach*. In Proc. of NLPIR4DL, (p. 88-95).

Kaplan, D., Tokunaga, T., and Teufel, S. 2016. *Citation Block Determination Using Textual Coherence*. Information Processing, 24(3), (p. 540-553).

Knoth, P., Gooch, P. and Jack, K. 2017. *What Others Say About This Work? Scalable Extraction of Citation Contexts from Research Papers*. In Proc. of TPDL, (p. 287299).

Mei, Q., and Zhai, C. 2008. *Generating Impact-Based Summaries for Scientific Literature*. In Proc. of ACL-HLT, (p. 816-824).

MacRoberts, M.H., and MacRoberts, B.R. 1984. *The Negational Reference: or the Art of Dissembling*. Social Studies of Science, 14, (p. 91-94).

Nakov, P., Schwartz, A., and Hearst, M. 2004. *Citances: Citation Sentences for Semantic Analysis of Bioscience Text*. In Proc. of SIGIR.

Nanba, H., and Okumura, M. 1999. *Towards Multi-paper Summarization Using Reference Information*. In Proc. of IJCAI, (p. 926-931).

O'Connor, J. 1982. *Citing Statements: Computer Recognition and Use to Improve Retrieval*. Information Processing and Management, 18(3), (p. 125-131).

Platt J.C. 1990. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. Advances in Large Margin Classifiers, (p. 61-74).

Qazvinian, V., and Radev, D. 2010. *Identifying Non-explicit Citing Sentences for Citation-based Summarization*. In Proc. of ACL, (p. 555-564).

Ritchie, A., Robertson, S., and Teufel, S. 2008. *Comparing Citation Contexts for Information Retrieval*. In Proc. of ACM-CIKM, (p. 213-222).

Sula, C., and Miller, M. 2014. *Citations, Contexts, and Humanistic Discourse: Toward Automatic Extraction and Classification*. Literary and Linguistic Computing, 29, (p. 453-464).

Teufel, S., Siddharthan, A., and Tidhar, D. 2006. *Automatic Classification of Citation Function*. In Proc. of EMNLP, (p. 103-110).