Natural Language Engineering **22** (3): 327–349. © Cambridge University Press 2015
doi:10.1017/S1351324915000388

Survey about citation context analysis: Tasks, techniques, and resources

MYRIAM HERNÁNDEZ-ALVAREZ[1] and
JOSÉ M. GOMEZ[2]

[1]*Escuela Politécnica Nacional, Facultad de Ingeniería de Sistemas, Quito, Ecuador*
e-mail: myriam.hernandez@epn.edu.ec
[2]*Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, España*
e-mail: jmgomez@ua.es

(*Received 6 October 2014; revised 8 September 2015; accepted 8 September 2015; first published online 5 November 2015*)

Abstract

Bibliometric calculations currently used to assess the quality of researchers, articles, and scientific journals have serious structural problems; many authors have noted the weakness of citation counts, because they are purely quantitative and do not differentiate between high- and low-citing papers. If a paper's reputation is simply evaluated according to the number of its citations, then incomplete, incorrect, or controversial articles may be promoted, regardless of their relevancy. Therefore, perverse incentives are generated for researchers who may publish many incorrect or incomplete papers to achieve high impact indexes. It is essential to improve the objective criteria for automatic article-quality assessments. However, to obtain these new criteria, it is necessary to advance the programmed detection of context, polarity, and function of bibliographic references.

We present an overview of general concepts and review contributions to the solutions to problems related to these issues, with the purpose of identifying trends and suggesting possible future research directions.

1 Introduction

Citation analysis is a method of evaluating the impact of an author, a published work, or scientific media. Sugiyama *et al.* (2010) suggested two types of citation analyses: citation counts (Garfield 1972) and citation context analysis. We accept this categorization, as it is close to our thesis.

Currently, scientific literature is mostly available on the Web, where automatic citation indexes are combined with online searches using tools, like CiteSeerX,[1] Google Scholar,[2] and Microsoft Academic Research.[3] Indexes obtained by counting citations provide valuable information on the behavior of papers, such as impact factor and H-index (Hirsch 2005). The impact factor measures citation frequency

[1] http://citeseerx.ist.psu.edu/
[2] http://scholar.google.com/
[3] http://academic.research.microsoft.com/

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

for a journal in a particular year, while the H-index is an indicator of an author's reputation, evaluating his/her scientific production depending on the citation count. Small (1973) proposed co-citation analysis to add a similarity measure between works A and B by counting the number of documents that cite them. This approach is complementary to the concept introduced by Kessler (1963) regarding 'bibliographic coupling,' which states that documents that have similar references deal with similar subjects.

Many authors have noted the weakness of citation count, because it is purely quantitative and does not differentiate between high- and low-citing papers. PageRank (Page, Brin, Motwani and Winograd, 1999) partially solved this problem with a rating algorithm that takes into account not only the counting of cited mentions but also the influence of citing journals. Sayyadi and Getoor (2009) went a step further and proposed FutureRank, which is the expected future PageRank score, based on upcoming citations, with a procedure that tried to predict future reference counting, using authorship network and the publication time of the article. In the same field of Citation Count Prediction (CCP), Yan *et al.* (2011) suggested paper topics and author expertise among other features as input for several regression models to foretell the estimated citation counts of that article after a given period of time. Davletov, Aydin and Cakmak (2014) dealt with CCP using citation count of the article when it was just published, modeling it as a regression problem.

Mei and Zhai (2008) extracted impact summaries, defining impact as the influence of the paper in topics related to its research subject, as reflected in its references.

Citation count analysis has valuable application in the development of bibliometric measures. Tools, such as Science Citation Index,[4] or Journal Citation Reports,[5] use impact factor to rank scientific journals. These indexes are useful and have had a profound impact on science management. However, there are controversies regarding their simple approaches, which consider all citations as equal regardless of the purpose or the polarity with which they were mentioned.

Radicchi (2012) showed that incomplete, erroneous, or controversial papers have higher citation counts. This can generate perverse incentives for new researchers who may be tempted to publish incomplete or erroneous results as a means of receiving a higher number of citations (Marder, Kettenmann and Grillner 2010). In fact, this affects the quality of prestigious journals, because it is known that accepting controversial articles is profitable and increases the overall citation numbers. Reviews, such as those conducted by the recently awarded Nobel Prize winner (Sample 2013), emphasize this issue. In this context, Brazilian journals have used self-references to skew the Journal Citation Reports index (Van Noorden 2013).

Another limitation of quantitative citation analysis is that all references are interpreted as if an author were influenced by the work of another, without specifying the type of influence (Zhang, Ding and Milojević 2013); this approach can omit the true impact of a citation (Ioannidis 2005; Young, Ioannidis and Al-Ubay 2008; Marder *et al.* 2010; Nicholson and Ioannidis 2012; Radicchi 2012; Brembs and

[4] http://scientific.thomson.com/products/jcr/
[5] http://purl.org/spar/cito

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Munafò 2013; Schreiber 2013). To understand the influence of a scientific paper, it is advisable to account for other, more qualitative criteria, such as an author's disposition towards a cited paper or the function of a citation in a paper. Not all citations are the same; for instance, a criticized quoted work does not have the same impact as a citation used as a conceptual starting point for a research paper. In the same sense, Reyhani *et al.* (2013) asserted that the citation number obtained by a paper is not a sufficient indicator to understand its influence; it is more important to know their contribution to citing papers. So, they proposed to calculate influence using similarity measures between citing and cited articles.

These problems add to the increasing importance of impact indexes in researchers' careers (Siegel and Baveye 2010). Pressure to publish seems to be the cause of increased fraud in scientific literature (Fang, Steen and Casadevall 2012). For these reasons, it is becoming more important to correct these problems and identify more complete metrics to evaluate researchers' relevance by taking into account other 'quality' factors related to citation context analysis. Citation context analysis seeks to find a contextual relationship between citing papers and cited papers using different methods that include natural language processing techniques, machine-learning systems, annotated datasets, statistical approaches, and dictionary approaches.

Citation context analysis applied to bibliometrics is reaching a point from which it will be possible to glimpse significant progress that could be applied to produce a more accurate index that includes qualitative and quantitative aspects for evaluating the relevance of science publications and researchers' work. For this type of assessment, additional information, such as the author's intention and disposition towards the cited work, is required. These criteria are related to citation functions and polarity. Automatic summarization and survey generation of scientific text and more complete citation indexers for web browsing are also of interest. Hence, it is important to conduct a detailed study of the current state of this issue to guide future research. Therefore, we evaluate the recent developments in the most researched areas in citation context analysis: citation context identification, citation functions, and polarity classification with their corresponding schemes and datasets.

Citation context identification is an important task in citation analysis. This approach is related to the identification of all parts of the text that are associated with a specific citation. Athar (2014) shows that integrating citation context into polarity citation analysis considerably increases the performance of citation classification algorithms.

According to Green *et al.* (2014), 'argumentation mining is a relatively new challenge in corpus-based discourse analysis that involves automatically identifying argumentative structures within a document—e.g., the premises, conclusion, and argumentation scheme of each argument…' Teufel (1999) and Teufel (2010) studied argumentative classification, including citation context. We think that argumentation mining could serve as a method to delimitate citation context, separating text that has arguments about a citation from text that is not related. It would be insightful to conduct studies with this focus, in which we would treat context definition as a topic linked to argumentation mining. Thus, we suggest that, as another approach, the length of the citation context could be established using argumentation mining

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

algorithms to automatically detect the argument around the citation. Nevertheless, despite recent advances in discourse parsing and causality detection, the automatic recognition of argumentation structure of texts is still a very challenging task (Peldszus 2014).

The context is also related to the structure of discourse in the paper. Teufel (2010) and Athar (2014) state that accounting for the location of the citation in the article would aid the classification of the citation's function. Athar (2014) suggests that it is likely that the most relevant citations are referenced several times in the paper, particularly in the middle of the main body. This feature captures the number of citations in an article, and it is an important characteristic for modeling the influence of a cited article within the citing work.

Citation functions are associated with the task of the reference in the citing paper. Citation function classification uses different schemes and categories for different citation roles; note that there is no consensus on a single scheme. These patterns are used to annotate corpus taken from scientific literature to train models for automatic citation classification.

Citation polarity is a more coarse-grained analysis that captures positive (favorable) or negative (unfavorable) dispositions towards a cited paper. Sometimes, polarity can be mapped to a citation function.

In general, citation context analysis addresses special types of documents in which it is difficult to detect the function of the reference or the author's disposition towards a quote because:

- Authors do not always explicitly state the citation's purpose.
- Authors avoid explicit criticism, particularly when it cannot be quantitatively justified. They often accomplish criticism through implications and hedging, defined as cautious language expressed in a vague form (Hyland 1996). MacRoberts and MacRoberts (1984) maintained that showing negative sentiment in a citation text is politically dangerous, and researchers prefer to disguise criticism toward the paper by praising its strengths in advance of a negative review if any.
- Many citations are simply mentioned as work that has been performed in the field. Ziman (1987) stated that many papers are cited out of 'politeness, policy, or piety.' They are passing references that are not really influential to the citing paper.
- There is a specific sentiment-related science lexicon that is primarily a compound of technical terms that diverge among scientific fields (Verlic *et al.* 2008). Often, academic expressions are specific to the knowledge in a field. Efforts to connect sentiments to technical terms and to develop general sentiment lexicons for science are underway (Small 2011).

For an integral methodology of citation context analysis, it is important to differentiate among references according to citation function and polarity, and from there, using additional information, such as location and citation frequency, it is possible to obtain an assessment of a cited document's influence on a citing paper. A holistic impact index should include not only citation counts but also an

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

individual valuation of influence; to achieve this, it will be necessary to overcome the aforementioned challenges. We divide this paper into sections to review work regarding these topics and their related issues: Citation Context Identification, Citation Function Classification, Citation Polarity Classification, and the corpora needed for research.

2 Citation context identification

Citation context is defined as the text surrounding references in scientific literature that is related to them. Context in which a citation appears provides valuable data about it (Kataria, Mitra and Bhatia 2010). The following are some approaches related to citation context identification.

2.1 Optimal size of context windows

Several works have identified the optimal size of context windows for detecting the sentences referring to the citation. Ritchie, Robertson and Teufel (2008) explored how to select context length to improve information retrieval of specific data. They experimented with variable window sizes, ranging from one sentence that included the citation to the entire text using linguistic motivations. They obtained better results with three sentences than with one and stated that simple windows are more effective that ones with full sentences.

Kataria *et al.* (2011) selected an adaptive window for context around a citation that statistically provides more information. They found window size through maximizing an objective function representing topical similarity between a cited document and its corresponding context, where topics are semantically significant clusters of words presented as co-ocurrences (Nallapati *et al.* 2008).

He, Kifer *et al.* (2011) analyzed methods to find citation context to make automatic recommendations to locate references. They used language models with 100-word overlapping windows, each one with a context probability score. The score sequence identifies high-probability regions that could be defined as citation context. They use contextual similarity around different parts of the document with topic relevance, given citation context clusters and a probabilistic model to compute relevance scores.

Athar and Teufel (2012) worked on a supervised approach with context windows of different lengths. They developed a classifier using citations as features sets with support vector machine (SVM) and uni-grams to tri-grams and dependency triplets as features. The authors explored the effect of context window lengths and obtained the best performance for citation classification using an annotated context window of four sentences. They showed that ignoring the citation context would result in a classification with more non-identified polarity due to loss of information, especially the criticism towards a cited paper. They recognize the need for better algorithms to filter contextual sentences.

Recommendation systems use context detection techniques that can be applied in this domain. Using citation context, they identify relevant papers that may

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

lead to specific recommendations of sections that need references (Livne *et al.* 2014).

As suggested before, we advocate for another approach in which detection of context should be related to the identification of the argument around the citation; thus, the context should be connected with the argumentation mining field. Manual processing or algorithms to automatically detect arguments around a citation could define its context. These arguments could be refuted or supported to infer different citation functions and polarity.

However, there are other views on this issue. As an example of unsupervised methods for citation context identification, we refer to the work of Kang and Kim (2012). Their corpus consisted of 56 articles from different academic journals that broadly belong to science and engineering fields; 1,048 citing sentences were yielded that used citation pattern rules with grammar presented in the Backus–Naur Form to detect citation contexts as phrases, clauses, sentences, multi-sentences, and other characteristics (e.g., figures, equations, and tables). Their manual experiments showed that multi-sentence citation context was only detected in 5% of the total citing sentences.

This result differs from most, in that the majority of experimental results suggests that to detect the polarity of references, the context size should be at least more than one statement (Jochim and Schütze 2012). The authors did not automatize the detection of citation contexts, but they define rules and suggest future work for using the contexts in a machine-learning model.

Athar and Teufel (2012) focused on features to automatically detect implicit reference mentions in citation contexts and showed that their inclusion improves the quality of classification. Features analyzed were the use of acronyms, cue phrases that signal continuation of the topics related to the citations, and use of third-person pronouns, among others. The resulting macro-average F-score was 0.754 for detecting implicit citations and 0.687 for polarity recognition.

Abu-Jbara and Radev (2012) addressed the problem of identifying fragments of a citing sentence that are related to a target reference–i.e., the reference scope. Their intent was to identify which segment refers to a given citation in a sentence with multiple citations. They approach context identification differently, because they considered a citing sentence as one that cites multiple papers.

To date, little is known about different contributions of each co-citation in an article. The authors compared three methods: word classification with SVM, sequence-labeling CRF, and segment classification. CRF-based sequence-labeling performed significantly better than the word classification method. Their results also showed that segment labeling performs better than word labeling. They used a corpus formed by 3,500 sentences that each contained at least two references; the material was randomly selected from papers found in the ACL Anthology Network corpus (ACL AAN)[6] (Bird *et al.* 2008). However, this work only refers to one citing sentence, rather than a broader context.

[6] Released Dec. 2013 http://clair.eecs.umich.edu/aan/index.php

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

2.2 Classification of contextual/non-contextual

Qazvinian and Radev (2010) classified sentences as contextual/non-contextual using three sentence-level features (similarity to reference, explicit citation, matches particular regular expressions) to train Markov Random Field and SVM models. The authors found the best results in a context window comprising four sentences before and after the reference. Note that their dataset consisted of 10 papers. Their F-macro value measure for the recognition of two classes, context and no context, was 0.87.

Context size delimitation was also conducted by Small (2011), who manually processed 20 papers of the citation summary data from the Association for Computational Linguistics Anthology Network (ACL AAN). They extract citation and co-citation contexts using a 'bag' of sample cue words to characterize dispositions towards cited work in maps of science. In their work, the context consisted of an average of 1.6 sentences around the point of reference.

A relationship exists between the location of a citation and its polarity and its importance within the article cited. Thus, it is necessary to take into account an established structure that normally has scientific papers—for example, IMRaD from Biber and Finegan (1994). Liakata *et al.* (2012) generated a tool with SVM and conditional random fields to train and compare 256 articles; 11 categories were recognized using Core Scientific Concepts (CoreSCs) to define context using text locations in various paper sections: Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result, and Conclusion. As discussed below, their results should be improved upon.

Another supervised technique is conditional random fields (CRFs) to tag, segment, and extract information from documents. The technique has been applied by some authors to identify context. Angrosh, Cranefield and Stanger (2013) manually annotated text in 20 articles as a development dataset and implemented CRF, to identify citation contexts. They established a model with seven context types for citation sentences and six no-context sentences. They applied the technique to three datasets, where each was composed of 10 articles; they obtained the best results for citation context identification, with an F1 value of 89.5%.

2.3 Limitations of experimental methods

Current research is rather limited to experimental methods to find optimal fixed-sized windows of context. Most of the current efforts use supervised learning algorithms. Unsupervised methods are less commonly used to define context, mainly because of the complexity of the task, which translates to a large set of rules that are difficult to handle; these methods are probably incomplete, because they cannot cover all possible cases.

An important limitation of this approach is the size of the corpora used by qualitative researchers. These researchers tend to use relatively small datasets, because generation of a specialized citation corpus is difficult and requires careful reading, professional knowledge, and expert judgment (Zhang *et al.* 2013). Automatic

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

annotations for generating large labeled corpus have generally produced poor accuracy.

Teufel, Siddharthan and Tidhar (2006) worked on a 2,829-sentence citation corpus from 116 articles using a 12-class classification scheme by applying context in annotations to determine an author's purpose for citing a text. The process was manual and based on rhetorical information. The authors established that sentences containing the citation often do not contain information about the relationship among the citing and cited papers; thus, it is necessary to define contextual sentences before and/or after the citation. Notably, in the context, frequent citations are implicitly named.

The results attained with the different approaches of context citation detection vary considerably. However, the results are not actually comparable due to the diversity of the used measures and corpora.

The most common measures are macro and micro F-scores (Qazvinian and Radev 2008; Athar and Teufel 2012; Kang and Kim 2012) and average precision (Abu-Jbara and Radev 2012; Athar and Teufel 2012; Angrosh *et al.* 2013). Small (2011) used a log-likelihood method to measure probability and find cue words in groups that expressed dispositions towards cited work.

Moreover, the use of different corpora is a large problem, because even though different measures can provide a slight indication of which systems are better, the use of incompatible corpora makes formal comparisons very difficult (Hernández and Gomez 2014). Standard resources are not available, and most of the corpora created or used in different studies have not been published.

There are some problems still to be resolved in the definition of citation context length. Automatizing detection of variable context length is a complex and yet incomplete task. In this field, context detection techniques could be applied for automatic recommendation systems or argumentation mining methods. However, until these approaches are fully developed, it is useful to apply a fixed number of sentences, as defined by experimental results.

3 Citation function classification

A citation's purpose reflects the author's intention when he added the reference. A citation function is related to the task of the reference in the citing paper. Here, the terms are often interchangeable.

Most citation classification approaches are not general applications but are rather heavily oriented towards specific science domains (Dong and Schäfer 2011). Almost all the research in this field corresponds to computer science and language technologies, while some research is associated with life science and biomedical fields.

Several authors have reviewed the contributions of the extraction and classification of citation functions in papers. Their objective was to identify citation functions or to detect if a citation was somewhat influential in the citing article (Teufel *et al.* 2006; Sugiyama *et al.* 2010; Dong and Schäfer 2011; Abu-Jbara and Radev 2012; Zhu *et al.* 2014) by considering not only that a reference is intended to acknowledge the

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

influence of other authors but also it has non-scientific motivations. It is important to differentiate one citation function from another to define which references are relevant to evaluate impact indexes.

Studies have focused on three issues: manual annotation of a corpus using different category and function schemes with different approaches; automatic labeling of the corpus from training data; and addressing the problem of defining features that will achieve the best results in citation classification. However, some results of the various studies should be improved upon.

3.1 Citation function categories

Additionally, it is important to define categories of citation function. Many proposals require human classification to differentiate citation function classes based on several types of criteria. A category review for citation function classification by several authors is shown in Table 1. Different annotation approaches present diverse levels of granularity in citation function definitions. These schemes define from 3 to 35 different classes. Less granularity often refers to polarity (positive, negative, or neutral/objective), as we will discuss later. Garzone (1997) and Garzone and Mercer (2000) proposed a 35-category scheme. Teufel (1999) established a basic scheme with three categories and a full scheme with four more. Teufel, Siddharthan and Tidhar (2009) defined a 12-category pattern that also maps polarity and function. Most citations are objective/neutral, even though Teufel *et al.* (2006) claim that making a reference is itself an act of acknowledgment. In general, fine-grained categories are suitable for applications, such as summary extraction and information retrieval with qualified citation indexing.

Another interesting issue related to citation function categorization is hedging. Mercer, DiMarco and Kroon (2004) conducted an introductory study that only applied little citation context data as hedging cues that were studied by Hyland (1998) to detect the dissembled intent of citations. In Mercer *et al.* (2004), frequency analysis was used to demonstrate that hedging cues tend to occur within the citation context and may help determine the purpose of a citation. Experimental work in this field has yet to be conducted.

3.2 Algorithms, tools, and results for function classification

Teufel *et al.* (2006) used cue words, part-of-speech (POS) tags, and different length n-grams to train a WEKA[7] machine learner IBk algorithm, and they obtained a macro F of 0.57 for replicating a human annotation for function classification, which is a difficult task. In the same experiment, their results for polarity classification presented a macro F of 0.71. As Teufel noted, annotation is reliable at 0.72.

Dong and Schäfer (2011) tested textual, physical (location, number of references, etc.), and syntactical features in 120 articles from the Association for Computational

[7] http://www.cs.waikato.ac.nz/ml/weka/

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Table 1. *Review of categories for citation functions and polarity classifications*

Author(s)	No. of categories	Categories for the annotation scheme
Garzone (1997)	35	7 negational; 5 affirmational; 4 assumptive; 1 tentative; 5 methodological; 3 interpretational/developmental; 1 future research; 2 conceptual; 2 contrastive; 5 reader alert.
Teufel (1999)	7	Basic scheme: background, other, own. Full Scheme: basic scheme + aim, textual, contrast, basis.
Teufel *et al.* (2009)	12	Negative polarity: weakness of cited approach; unfavorable contrast/comparison. Neutral polarity: contrast/comparison in goals of methods; contrast/comparison in results; contrast between two cited methods; neutral description of cited work, or not enough textual evidence for above categories or unlisted citation function. Positive polarity: author uses cited work as starting point; author uses tools/algorithm/data; author adapts or modifies tools/algorithm/data; this citation is positive regarding approach or problem addressed; author's work and cited work are similar; author's work and cited work are compatible/support each other.
Dong and Schäfer (2011)	3 polarity 4 function	Polarity: positive, negative and neutral. One dimension of Moravcsik and Murugesan, organic versus perfunctory divide into background, fundamental idea, technical basis, comparison.
Liakata *et al.* (2012)	11	Core scientific concepts: hypothesis, motivation, goal, object, background, method, experiment, model, observation, result and conclusion.
Jochim and Schütze (2012)	4 functions (one of them corresponds to polarity)	Apply Moravcsik and Murugesan scheme with four orthogonal facets that determined whether the reference is an idea or a tool, organic or perfunctory, building upon the cited work or presenting an alternative, confirmative or negational (this category could correspond to polarity: positive or negative).
Angrosh *et al.* (2013)	7 Non-citation; 7 Citation Sentences.	Non-citation sentences: background; issues; gaps; description; current work outcome; future work. Citation sentences: cited work identified gaps; cited work overcomes gaps; uses outputs from cited works; result with cited work; compare works of cited work; shortcoming in cited work; issue related cited work.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Table 1. *Continued*

Author(s)	No. of categories	Categories for the annotation scheme
Li *et al.* (2013)	3 polarity 12 function	Positive: based upon, corroboration, discover, positive, practical, significant, standard, and supply. Neutral: contrast, co-citation, neutral. One negative function.
Ciancarini *et al.* (2013)	13	Agrees with, cites, cites as author, cites as authority, cites a data source, cites as evidence, cites as metadata document, cites as potential solution, cites as recommended reading, cites as related, confirms, corrects, critiques, derides.
Iorio *et al.* (2013)	3 polarity 7 function	Polarity: positive, negative, neutral. Function: citesfor information, usesMethodIn, usesConclussionsFrom, obtainsBackgroundFrom, discusses, citesAsRelated and others, credits and others.
Meyers (2013)	2	Corroborate and contrast

Linguistics Anthology Reference Corpus (ACL ARC)[8] (Radev, Muthukrishnan and Qazvinian 2009) as training data. The results produced F1 macro values of 0.66 for citation function classification (fundamental idea/technical basis/comparison). They used a semi-supervised automatic data annotation with an ensemble-style self-training algorithm in approximately 170 training instances. The context was as many as three sentences.

Jochim and Schütze (2012) used the model of Moravcsik and Murugesan (1975) with four facets to establish whether cited work is an idea or a tool, whether it is accurate or faulty, whether it is fundamental or perfunctory, and whether the citing paper builds upon or is an alternative to the cited paper. The authors stated that each of the Moravcsik and Murugesan facets can be used according to the type of classification needed. The facet that established whether the work was accurate or faulty (confirmatory/negational) can be mapped directly to polarity. The Stanford Maximum Entropy classifier (Klein and Manning 2003) was applied, in which context lengths of 1, 2, and 3 sentences are identified. The best results were obtained for sentence lengths greater than one. The authors reported a good feature analysis for each facet and offered their annotated corpus to the scientific community (Table 2). In their work, they used a combination of features: unigrams, word-level linguistic features, comparatives, sentence location, and a lexicon with positive and negative sentences. Their F1 scores varied as follows: 68.2 (discriminating between an idea and a tool); 51.1–52.9 (work that evolved from or was an alternative to a cited paper); 58.0 for organic (fundamental) or perfunctory research; and 51.1 for confirmational/negational facets.

[8] http://acl-arc.comp.nus.edu.sg/

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Table 2. *Available annotated corpora for citation polarity classification*

Corpus	Number of citations	Positive	Negative	Neutral/objective
DFKI	1,768	190	57	1,521
IMS	2,008	1,836	172	–
CITD	8,736	829	280	7,627

Meyers (2013) classified two functions: corroborate and contrast. Corroborate describes two works that follow the same approach, while contrast describes two differing approaches or opinions. The author used a lexical dictionary and manually written rules to create an algorithm based on tree-modeled discourse. The results were acceptable—i.e., 216 correct answers for 291 citation relations were obtained. The algorithm was applied to 20 PubMed[9] scientific articles, with recall values of 67% for contrasting and 83% for corroborating papers. These results need to be proven in a larger corpus.

Ciancarini *et al.* (2013) annotated a corpus for citation functions using citation contexts and 13 function categories. They did not report the experimental results of their corpus, but they recognized that their work is preliminary.

Using a similar approach, Iorio, Di, Nuzzolese and Peroni (2013) implemented a citation-extraction algorithm through pattern matching. They detected polarity using CiTalO, a software developed using a combination of techniques of ontology learning from natural language, sentiment analysis, word-sense disambiguation (with SVM), and ontological mapping. The corpus was 18 papers published in the seventh volume of the Balisage Proceedings[10] (2011) regarding Web mark-up technologies. With the highest recall of only 0.491, the results need improvement. CiTalO implements properties of CiTO[11] ontology.

Li *et al.* (2013) used a maximum entropy-based system to create a training set from annotated data and automatically classify functions into two levels of granularity; three polarity categories were linked to the sentiment of 12 citation functions. The authors used 91 annotated articles with 6,355 citation instances. The best F1 performance was 0.67.

Abu-Jbara, Ezra and Radev (2013) used a trained classification model and SVM classifier with a linear kernel and AAN dataset. They selected 30 papers that received a total of 3,500 citations within AAN in 1,493 papers. For development data, they created another dataset with 300 citations that cite five papers from AAN to determine the optimal citation context window. The purpose (function) classification results had a macro F of 58%. For automatic classification, the accuracy would probably increase as more training data become available (Abu-Jbara *et al.* 2013).

[9] http://www.ncbi.nlm.nih.gov/pubmed
[10] http://balisage.net/Proceedings/vol7/cover.html
[11] http://vocab.ox.ac.uk/cito

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

In general, the best results for citation function or purpose classification were obtained using a SVM algorithm, with F1 values as high as 73.1, depending on the selected features.

3.3 Unresolved problems

The main unresolved problems in citation classification are the lack of a standard categorization scheme for functions and the development of a public access golden corpus with sufficient features to perform citation classification that yields good results in precision, recall, and F-measures. Currently, low values for these results are obtained, because for most schemes, their training annotated corpus is small and does not have the features needed to provide sufficient information to the classifier algorithms for their task.

4 Citation polarity classification

Polarity is defined as a coarse-grained approach to an author's sentiment towards a citation. It can be defined as positive, negative, or neutral/objective. This type of citation classification is closely related to sentiment analysis.

Many algorithmic approaches can be defined for these tasks; machine-learning systems have proven to be effective in many text classification tasks, particularly in the sentiment analysis area (Zhang, Yu and Meng 2007). A comparative study was performed by Sebastiani (2002) that concluded that the best classifier applied to sentiment analysis was the super vector machine (SVM). Subsequent studies have confirmed these findings (Mullen and Collier 2004; Pang and Lee 2004; Wilson *et al.* 2005; Boldrini *et al.* 2009; Prabowo and Thelwall 2009; Fernández *et al.* 2011). These studies have been performed using general texts or have been extracted from the contents of Web 2.0, such as reviews, user comments, blogs, and microblogs; their extrapolation to the domain of citations in scientific literature is still debated.

Polarity classification could be useful for improving bibliometric measures to evaluate the actual citation impact of a cited work in a citing paper. Citation polarity could be an additional criterion to improve measurements of a paper's impact by considering the differences between a citation that is positive from a citation that is criticized or refuted, which should have less weight or negative weight in the calculation of impact indexes (Abu-Jbara *et al.* 2013).

To classify polarity citations, different feature sets can be used with varying degrees of effectiveness. As we mentioned, in general, results from different studies are not comparable because of the diversity of corpora, but we may begin to assess the effects of different features in citation polarity classification using authors' results. Hedging should also be considered, because it is used to disguise an author's negative disposition towards a cited work, but it is not yet used for this task.

Jochim and Schütze (2012) collected and classified different types of features used in previous work and evaluated their effects on the classification of different facets. For polarity classification, we are particularly interested in the confirmational/

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

negational facet of the Moravcsik and Murugesan scheme, because it may be mapped to positive and negative polarity. The authors detected that for this facet, a Bag-of-Word (BoW) classified as positive or negative produced better results than other features for identifying polarity, along with word-level linguistic features, such as modal auxiliary 'has' and main verbs, first- and third-person pronouns, comparatives, superlatives, and contrasting conjunctions. They also concluded that feature classes, such as lexical and frequency, are detrimental for classification of this facet.

In similar approaches, Teufel *et al.* (2006) used verb tense, voice, and so-called modal auxiliaries, such as 'may' and 'could,' the overall location of a sentence, and POS tagging. For machine learning, they used the K-nearest neighbors classifier—i.e., the IBk algorithm with k = 3 in the WEKA tool kit—to replicate human annotation of a corpus classification of weak, positive, and neutral categories (the macro F was 0.71). Athar (2011) used word-level linguistic features with POS tags to achieve polarity classification. Abu-Jbara *et al.* (2013) conducted a BoW analysis using lists of negation, subjectivity, speculation, and other similar cues to detect polarity.

Teufel *et al.* (2009), Dong and Schäfer (2011), and Jochim and Schütze (2012) showed a connection between citation functions and polarity in its location in a sentence, paragraph, or section in the paper. Teufel *et al.* (2006) emphasized the relationship among sections of the paper where a citation was located near its function and polarity.

For polarity citation classification, Athar (2014) used a corpus consisting of 852 papers that cite the top 20 target papers. Athar obtained the best results for the SVM classifier (macro F = 0.808, micro F = 0.764) over the Naive Bayes classifier (macro F = 0.471, micro F = 0.755) using 1–3 grams, dependency features, and window-based negation. Dependency structures are triplets, or relations (head, dependent) that capture the long-distance relationships between words. Window-based negation considers that all words inside a k-word window (k = 1–15) of any negation term are suffixed to distinguish them from non-polar versions. Categories for different schemes for polarity classification and citation function are shown in Table 1.

Athar (2014) reported fewer citation occurrences with negative polarity. Some of these negative mentions cannot be detected, because authors tend to hide negative disposition to cited papers through the use of hedging. Effective hedging detection is a necessary task, for which linguistic clues need to be considered, as detailed by Hyland (1996). Techniques should be used to address more abundant neutral citations that can skew classification results.

5 Corpora

The most widely used corpora are subsets obtained from the ACL Anthology, which is a digital archive of research papers in computational linguistics. These subsets are the ACL ARC (Radev, *et al.*, 2009) and the Association for Computational Linguistics Anthology Network (ACL AAN) (Bird *et al.*, 2008); they serve as the basis for a number of experiments that use parts of these large bodies of data.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Other corpora are annotated subsets from the Linguistic Data Consortium (LDC)[12]; this version adds to the ACL ARC, with several page images in both text and image forms. PubMed and the Lecture Notes in the Computer Science (LNCS)[13] series published by Springer are also used as basic corpora.

5.1 General corpora

The datasets acquired from subsets of the aforementioned large datasets were manually labeled according to the different approaches of the various tasks. The Association for Computational Linguistics Anthology Network (ACL AAN) was used to select papers for annotation and study by Athar and Teufel (2012); Abu-Jbara *et al.* (2013); and Tsai, Kundu and Roth (2013). The ACL ARC was used as a basic dataset by Sugiyama *et al.* (2010), Athar (2011), and Abu-Jbara and Radev (2012). Some articles were chosen from PubMed by Li *et al.* (2013) and Meyers (2013). Additionally, some authors work with collections obtained from LNCS (Kaplan, Iida and Tokonuga 2009; Angrosh *et al.* 2013).

Prior to the formation of the Association for Computational Linguistics Anthology Network (ACL AAN), researchers, such as Teufel and Moens (1999), chose papers presented at conferences that later became part of that archive.

The corpora mentioned do not seem to be available to the research community. Human annotation is often evaluated using an inter-annotator agreement kappa based on the difference between observed and expected agreements.

5.2 Corpora with function and polarity annotations

We found only three public datasets containing citations with annotations of functions and polarity. The first dataset was DFKI citation corpus[14] (Dong and Schäfer 2011), which takes subsets of the articles of the ACL Proceedings of 2007 and 2008.

The second dataset is IMS Citation Corpus[15] (Jochim and Schütze 2012), which uses articles from the ACL Proceedings of 2004. Both datasets are taken from the ACL ARC corpus. The DFKI Citation Corpus includes polarity annotations. The authors used the Moravcsik and Murugesan scheme, where positive entries correspond to confirmative and negative entries are negational citations. Jochim (2014) compiled these two corpora into CITD, which integrated unlabeled data from other years of ACL Proceedings—i.e., 1979–2003, 2005–2006, and 2009—with the Multi-Domain Sentiment Dataset MDSD Corpus[16] (Blitzer, Dredze and Pereira 2007). This corpus compiles product reviews from Amazon.com for classifying quotes by applying the marginalized, stacked, denoising auto-encoders mSDA

[12] https://www.ldc.upenn.edu/
[13] http://www.springer.com/computer/lncs?SGWID = 0-164-0-0-0
[14] https://aclbib.opendfki.de/repos/trunk/citation_classification_dataset/
[15] http://www.ims.uni-stuttgart.de/~jochimcs/citation-classification/
[16] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

algorithm introduced by Chen *et al.* (2012) with a linear SVM. In Table 2, we show the three previously mentioned annotated corpora for citation polarity classification.

5.3 Problems presented in corpus annotation

Manual annotation of a citation context analysis is a complex and time-consuming task. For this reason, annotated available corpora are small and not publicly available. The manual annotation process has problems, such as low inter-annotation agreement, due to the cognitive complexity of this job (Ciancarini *et al.* 2014). A good agreement is a requirement for reliability and reproducibility of the annotation schemes (Artstein and Poesio 2008). For a manually annotated corpus, Teufel *et al.* (2006) presented a good inter-annotator agreement of 0.79.

Automatic corpus annotation for citation context analysis has not had good outcomes. Some semi-supervised techniques have been applied to automatic labeling of corpus data. Liakata *et al.* (2012) obtained an automatically annotated corpus with citation context based on discourse structure with an accuracy of around 0.50 when classifying the 11 CoreSCs. They used a CRF algorithm. Additional classes are more difficult to classify, and these results should be improved before this type of corpus is ready for use.

In this area, there are tasks that are still to be improved, such as: the definition of a golden corpus scheme for citation context analysis; manual annotation according to this standard with adequate coded features to yield good precision in automatic classification; and the development of methods for programmed marking of a dataset based on the aforementioned golden corpus scheme.

6 Discussion and conclusions

The development of citation context-identification applications should be based on argumentative mining techniques; however, so far, the applications are mainly implemented as heuristic approaches. Detection of arguments around a citation—i.e., what is said about a cited work—should define the citation context. Most of the work for automatic context identification uses machine-learning algorithms with supervised approaches; preferred features are selected via Bag-of-Words, n-grams of length 1 to 3, and dependency triplets. Better results in automated context classification are achieved when fewer classes exist. For example, for two classes, context and no context results have accuracies as high as 0.89. For more classes, as in the case of the 11-category CoreSCs, the performance accuracy is only 0.50.

For corpus data generation, manual annotation is difficult and lengthy. Automatic annotation using semi-supervised and supervised techniques has produced poor results. As a consequence, there is no corpus available with sufficient size and reliability that is accessible to the scientific community. Thus, researchers work with small datasets that they create themselves, and they use these datasets for

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

manual experiments. Attempts to obtain robust automatic classifiers have not been successful.

The scope problem for a given sentence with multiple citations has been undertaken using a context of only one sentence to define what is said about each reference, but additional work is required to improve results.

Detection of implicit citations has had favorable outcomes, but techniques need to be tested in larger corpora.

Citation function classification aims to extract and categorize citation functions from scientific literature. One possible application is to detect if a citation is somewhat influential in a citing article. This type of information would be useful for improving citation indexes, because influential references should have more weight than general or perfunctory citations. Similarly, if citing articles are built upon material from a cited paper, then this reference should have more weight than criticized references.

Of course, a classification scheme must first be defined to classify functions. Different schemes have been proposed with different granularity (35 to 3 classes). There is no standard defined for citation-classification function schemes. Citation polarity classification is a coarse-level sentiment analysis that could be seen as a citation function classification with two or three classes: positive, negative, and/or neutral. Polarity classification also has bibliometric applications, because criticized citations should have less weight in the evaluation of the impact indexes.

Another important issue is hedging as a means to disguise an author's motivation for making a reference. Experimental work should develop recognition of hedging to enhance detection and classification of citation functions and polarity.

Feature engineering is an essential problem for function and polarity detection. It seems that feature selection has more impact than algorithm selection in citation classification. Depending on the classification scheme, different features become more appropriate and efficient. For citation function classification, features will depend on the categories defined. For polarity classification, it appears that the use of the location of the reference, semantic analysis at the word-level (positive and negative implications), and word-level linguistic characteristics are the most efficient features.

Therefore, in related sections of this paper, we elaborate that there are issues and interesting topics for the future research in these fields, such as:

- Detecting context windows of different lengths. We proposed that the window length should be related to argument mining such that the context includes most of what is said about a reference or that similarity measures be applied to detect relevant context around a citation.
- Identifying all references in the cited work, including those that are non-explicit, using natural-language-processing techniques and discourse analysis.
- Developing and applying domain-independent techniques and feature engineering for citation classification. Definition of annotated features is vital for achieving good classification results on precision, recall, and F-measure.
- Detecting hedging to distinguish uncertain events that have an impact in function and polarity classification.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

- Developing a unique framework to execute experimental comparisons among different techniques. This framework should cover available annotated corpora and standard citation categories.

In conclusion, in this paper, we propose a survey of context identification; function classification; polarity analysis of citations; and related research areas, such as feature engineering, schemes for categorization of citation functions and polarity, manual and automatic creation of corpora, application of experiments to define citation contexts, and argumentative mining techniques for defining contexts.

The major problem facing researchers is that there is no available corpus with sufficient size and standard annotation schemes that can be used in a shared form to facilitate results that are comparable.

Under these conditions, it is difficult to evaluate the efficiency of different approaches and techniques. Obviously, it is necessary to establish conditions that allow and motivate collaborative work to achieve tangible goals in these fields. The results could be translated into the development of applications that facilitate studies of the dynamics of scientific literature and that promote integral bibliometric measures.

Acknowledgements

This research has been partially funded by Universidad de Alicante, Spanish Government, Europe Union, and Valencian Community through the following projects: 'Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario' (GRE13–15), 'Tratamiento inteligente de la información para la ayuda a la toma de decisiones' (GRE12-44), ATTOS (TIN2012-38536-C03-03), LEGOLANG-UAGE (TIN2012-31224), SAM (FP7-611312), FIRST (FP7-287607), and DIIM2.0 (PROMETEOII/2014/001).

References

Abu-Jbara, A., and Radev, D. 2012. Reference scope identification in citing sentences. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL. Stroudsburg, PA, pp. 80–90.

Abu-Jbara, A., Ezra, J., and Radev, D. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL. Atlanta, GA, pp. 596–606.

Angrosh, M. A., Cranefield, S., and Stanger, N. 2013. Conditional random field based sentence context identification: Enhancing citation services for the research community. In *Proceedings of the First Australasian Web Conference*, Adelaide, Australia, Australian Computer Society, Inc., vol. **144**: pp. 59–68.

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* **34**(4): 555–96.

Athar, A. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, ACL. Stroudsburg, PA, pp. 81–7.

Athar, A. 2014. Sentiment analysis of scientific citations. Technical Report, University of Cambridge, Computer Laboratory, (UCAM-CL-TR-856).

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Athar, A., and Teufel, S. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL. Montreal, Canada, pp. 597–601.

Biber, D., and Finegan, E. 1994. Intra-textual variation within medical research articles. In N. Oostdijiik and P. DeHaan (eds.), *Corpus-Based Research into Language*, pp. 201–22. Amsterdam: Rodopi.

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M. Y., Lee, D., Powley, B., Radev, D. R., and Tan, Y. F. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the 6th International Conference on Language Resources and Evaluation Conference (LREC'08)*, Marrakesh, Morocco, pp. 1755–59.

Blitzer, J., Dredze, M., and Pereira, F. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 440–47.

Boldrini, E., Fernández Martínez, J., Gómez Soriano, J. M., and Martínez Barco, P. 2009. Machine learning techniques for automatic opinion detection in non-traditional textual genres. In *Proceedings of the First Workshop on Opinion Mining and Sentiment Analysis, WOMSA09*, Seville, Spain, pp. 110–19.

Brembs, B., and Munafò, M. 2013. Deep impact: Unintended consequences of journal rank. Digital Libraries; Physics and Society. Available at http://arxiv.org/abs/1301.3748

Chen, M., Xu, Z., Weinberger, K., and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, pp. 767–84.

Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., and Vitali, F. 2014. Evaluating citation functions in CiTO: Cognitive issues. *Semantic Web: Trends and Challenges*, pp. 580–94. Berlin: Springer International Publishing.

Ciancarini, P., Iorio, A. Di, Nuzzolese, A. G., Peroni, S., and Vitali, F. 2013. Semantic annotation of scholarly documents and citations. *AI*IA 2013: Advances in Artificial Intelligence*, vol. **8249**: pp. 336–47. Berlin: Springer.

Davletov, F., Aydin, A. S., and Cakmak, A. 2014. High impact academic paper prediction using temporal and topological features. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai, China, pp. 491–98.

Dong, C., and Schäfer, U. 2011. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing*, Chiang Mai, Thailand, pp. 623–31.

Fang, F. C., Steen, R. G., and Casadevall, A. 2012. Misconduct accounts for the majority of retracted scientific publications. In *Proceedings of the National Academy of Sciences of the United States of America*, United States of America, vol. **109**, pp. 17028–33.

Fernández, J., Boldrini, E., Gómez, J. M., and Martínez-Barco, P. 2011. Evaluating EmotiBlog robustness for sentiment analysis tasks. In *Natural Language Processing and Information Systems*, Heidelberg: Springer-Verlag, pp. 290–94.

Garfield, E. 1972. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science* **178**: 471–79.

Garzone, M. A. 1997. Automated classification of citations using linguistic semantic grammars. *Master's Thesis*. The University of Western Ontario. Available at http://www.collectionscanada.gc.ca/obj/s4/f2/dsk2/ftp04/mq28570.pdf

Garzone, M., and Mercer, R. E. 2000. Towards an automated citation classifier. In *Advances in Artificial Intelligence*, pp. 337–46. Berlin Heidelberg: Springer.

Green, A., Ashley, K., Litman D., Reed C., and Walker V. 2014. In *Proceedings of the First Workshop on Argumentation Mining*, ACL. Baltimore, MD, p. 3.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

He, Q., Kifer, D., Pei, J., Mitra, P., and Giles, C. L. 2011. Citation recommendation without author supervision. In *Proceedings of the 4th ACM international Conference on Web Search and Data Mining*, ACM. Kowloon, Hong Kong, pp. 755–64.

Hernández, M., and Gómez, J. M. 2014. Survey in sentiment, polarity and function analysis of citation. In *Proceedings of the First Workshop on Argumentation Mining*, ACL. Baltimore, MD, pp. 102–3.

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America,* United States of America, **102**(46): 16569–72.

Hyland, K. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics* **17**: 433–54.

Hyland, K. 1998. *Hedging in Scientific Research Articles*, vol. 54. Amsterdam: John Benjamins Publishing.

Ioannidis, J. P. A. 2005. Why most published research findings are false. *Chance* **18**(4): 40–7.

Iorio, A., Di, Nuzzolese, A. G., and Peroni, S. 2013. *Towards the Automatic Identification of the Nature of Citations*. Montpellier, France: SePublica, pp. 63–74.

Jochim, C. 2014. Improving citation polarity classification with product reviews. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL. Baltimore, MD, pp. 42–8.

Jochim, C., and Schütze, H. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Procedings of COLING'12*, Mumbai, India, pp. 1343–58.

Kang, I.-S., and Kim, B.-K. 2012. Characteristics of citation scopes: a preliminary study to detect citing sentences. In *Computer Applications for Database, Education, and Ubiquitous Computing Information Science*, pp. 80–5. Berlin: Springer.

Kaplan, D., Iida, R., and Tokunaga, T. 2009. Automatic extraction of citation contexts for research paper summarization: a coreference-chain based approach. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, ACL. Suntec, Singapore, pp. 88–95.

Kataria, S., Mitra, P., and Bhatia, S. 2010. Utilizing context in generative bayesian models for linked corpus. In *AAAI Conference in Artificial Intelligence*, Atlanta, Georgia, USA, pp. 1340–45.

Kataria, S., Mitra, P., Caragea, C., and Giles, C. L. 2011. Context sensitive topic models for author influence in document networks. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, Barcelona, Spain, vol. **22**(3): p. 2274.

Kessler, M. M. 1963. Bibliographic coupling between scientific papers. *American documentation Wiley Periodicals, Inc*. **14**(1): 10–25.

Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting*, ACL. Stroudsburg, PA, USA, vol. **1**: pp. 423—30.

Li, X., He, Y., Meyers, A., and Grishman, R. 2013. Towards fine-grained citation function classification. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 402–7.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* **28**: 991–1000.

Livne, A., Gokuladas, V., Teevan, J., Dumais, S. T., and Adar, E. 2014. CiteSight: supporting contextual citation recommendation using differential search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM. Gold Coast, Australia, pp. 807–16.

MacRoberts, M. H., and MacRoberts, B. R. 1984. The negational reference: Or the Art of dissembling. *Social Studies of Science*, London, Beverly Hills and New Delhi, **14**(1): 91–4, Sage Publications Ltd.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Marder, E., Kettenmann, H., and Grillner, S. 2010. Impacting our young. In *Proceedings of the National Academy of Sciences of the United States of America* United States of America, **107**: 21233.

Mei, Q., and Zhai, C. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of the 46 Annual Meeting: HLT*, ACL. Columbus, Ohio, USA, vol. **8**: pp. 816–24.

Mercer, R. E., Di Marco, C., and Kroon, F. W. 2004. The frequency of hedging cues in citation contexts in scientific writing. In *Advances in Artificial Intelligence*, pp. 75–88. Berlin: Springer Heidelberg.

Meyers, A. 2013. Contrasting and corroborating citations in journal articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, Hissar, Bulgaria, pp. 460–66.

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

Mullen, T., and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Conference on Empirical Methods in Natural Language Processing*, ACL. Barcelona, Spain, pp. 412–18.

Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM. Las Vegas, Nevada, USA, pp. 542–50.

Nicholson, J. M., and Ioannidis, J. P. A. 2012. Research grants: Conform and be funded. *Nature* **492**: 34–6.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab, Stanford University (SIDL-WP-1999–0120).

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL. Morristown, NJ, USA, pp. 271–78.

Peldszus, A. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining ACL 2014*, Baltimore, MD, USA, pp. 88—97.

Prabowo, R., and Thelwall, M. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics* **3**: 143–57.

Qazvinian, V., and Radev, D. R. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics*, ACL. Stroudsburg, PA, vol. 1: pp. 689—96.

Qazvinian, V., and Radev, D. R. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting*, ACL. Uppsala, Sweden, pp. 555—64.

Radev, D. R., Muthukrishnan, P., and Qazvinian, V. 2009. The ACL Anthology Network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, ACL. Suntec, Singapore, pp. 54–61.

Radicchi, F. 2012. In science "there is no bad publicity": Papers criticized in comments have high scientific impact. *Nature Scientific Reports* **2**: 815.

Reyhani Hamedani, M., Kim, S. W., Lee, S. C., and Kim, D. J. 2013. On exploiting content and citations together to compute similarity of scientific papers. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ACM. San Francisco, CA, USA, pp. 1553–56.

Ritchie, A., Robertson, S., and Teufel, S. 2008. Comparing citation contexts for information retrieval. In *Proceedings of the 17th Acm Conference on Information and Knowledge Management*, ACM. Napa Valley, CA, USA, pp. 213–22.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Sample, I. 2013. Nobel winner declares boycott of top science journals. *The Guardian*. Available at http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals

Sayyadi, H., and Getoor, L. 2009. FutureRank: Ranking scientific articles by predicting their future PageRank. In *SDM Siam International Conference on Data Mining*, Sparks, Nevada, pp. 533–44.

Schreiber, M. 2013. A case study of the arbitrariness of the h-index and the highly-cited-publications indicator. *Journal of Informetrics* **7**: 379–87.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* **34**: 1–47.

Siegel, D., and Baveye, P. 2010. Battling the paper glut. *Science* **329**: 1466.

Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* **24**: 265–69.

Small, H. 2011. Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics* **87**: 373–88.

Sugiyama, K., Kumar, T., Kan, M.-Y., and Tripathi, R. C. 2010. Identifying citing sentences in research papers using supervised learning. In *2010 International Conference on Information Retrieval and Knowledge Management (CAMP)*, Shah Alam, Selangor, Malaysia, pp. 67–72.

Teufel, S. 1999. Argumentative zoning: Information extraction from scientific text. Doctoral dissertation, School of Cognitive Science, University of Edinburgh, UK. Available at http://www.cl.cam.ac.uk/∼sht25/thesis/t1.pdf

Teufel, S. 2010. The structure of scientific articles: Applications to citation indexing and summarization. *CLSI –Studies in Computational Linguistics*, Chicago: University of Chicago Press.

Teufel, S., and Moens, M. 1999. Discourse-level argumentation in scientific articles: Human and automatic annotation. In *Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop*, ACL. Somerset, NJ, USA, pp 84–93.

Teufel, S., Siddharthan, A., and Tidhar, D. 2006 July. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ACL. Stroudsburg, PA, pp. 103–10.

Teufel, S., Siddharthan, A., and Tidhar, D. 2009. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, ACL. Stroudsburg, PA, pp. 80—7.

Tsai, C.-T., Kundu, G., and Roth, D. 2013. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, ACM Press. New York, NY, USA, pp. 1733–38.

Van Noorden, R. 2013. Brazilian citation scheme outed. *Nature* **500**(7464): 510–11.

Verlic, M., Stiglic, G., Kocbek, S., and Kokol, P. 2008. Sentiment in Science – A Case Study of CBMS Contributions in Years 2003 to 2007. In *21st IEEE International Symposium on Computer-Based Medical Systems*, University of Jyväskylä, Finland, pp. 138–43.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., and Patwardhan, S. 2005. OpinionFinder. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, ACL. Morristown, NJ, USA, pp. 34–5.

Yan, R., Tang, J., Liu, X., Shan, D., and Li, X. 2011. Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ACM. Glasgow, UK, pp. 1247–52.

Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. 2008. Why current publication practices may distort science. *PLoS Medicine* **5**(10): e201.

Zhang, G., Ding, Y., and Milojević, S. 2013. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology* **64**(7): 1490–503.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press

Zhang, W., Yu, C., and Meng, W. 2007. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*, ACM Press. New York, NY, USA, p. 831.

Zhu, X., Turney, P., Lemire, D., and Vellino, A. 2014. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology* **66**(2): 408–27.

Ziman, J. M. 1987. *An Introduction to Science Studies: The Philosophical and Social Aspects of Science and Technology*, Cambridge: Cambridge University Press.

https://doi.org/10.1017/S1351324915000388 Published online by Cambridge University Press