Natural Language Engineering **23** (1): 93–130. © Cambridge University Press 2016
doi:10.1017/S1351324915000443

NLP-driven citation analysis for scientometrics

RAHUL JHA[*1], AMJAD-ABU JBARA[*1], VAHED QAZVINIAN[2] and DRAGOMIR R. RADEV[3]

[1]*Microsoft Corp., Redmond, WA, USA*
e-mails: rajh@microsoft.com, amjada@microsoft.com
[2]*University of Michigan, Ann Arbor, MI, USA*
e-mail: vahed@umich.edu
[3]*EECS and SI, University of Michigan, Ann Arbor, MI, USA*
e-mail: radev@umich.edu

(*Received 18 January 2015; revised 17 November 2015; accepted 18 November 2015; first published online 25 January 2016*)

Abstract

This paper summarizes ongoing research in Natural-Language-Processing-driven citation analysis and describes experiments and motivating examples of how this work can be used to enhance traditional scientometrics analysis that is based on simply treating citations as a 'vote' from the citing paper to cited paper. In particular, we describe our dataset for citation polarity and citation purpose, present experimental results on the automatic detection of these indicators, and demonstrate the use of such annotations for studying research dynamics and scientific summarization. We also look at two complementary problems that show up in Natural-Language-Processing-driven citation analysis for a specific target paper. The first problem is extracting citation context, the implicit citation sentences that do not contain explicit anchors to the target paper. The second problem is extracting reference scope, the target relevant segment of a complicated citing sentence that cites multiple papers. We show how these tasks can be helpful in improving sentiment analysis and citation-based summarization.

1 Introduction

The field of scientometrics focuses on analyzing 'the quantitative aspects of the generation, propagation, and utilization of scientific information' (Braun, Bujdosó and Schubert 1987). Recent years have seen increased adoption of scientometrics techniques for assessing research impact of publications, researchers, institutions, and venues. For most scientometric measures, a citation is treated as the basic unit of impact (Vinkler 2010). A citation from the citing publication to the cited publication counts as a 'vote' for the impact of the paper and aggregate citation statistics are then used to come up with evaluative metrics for measuring scientific impact.

Despite several criticisms of citation-based measures (Waltman, van Eck and Wouters 2013; Bornmann and Marx 2014; Gorraiz, Gumpenberger and Schlögl

[*]This research was conducted while the authors were at University of Michigan.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

2014; Jonkers, Lopez-Illescas and Besselaar 2014) and proposals for using impact measured through other sources such as the Web (Brody, Harnad and Carr 2006; Thelwall *et al.* 2013) and Twitter (Haustein *et al.* 2014; Eysenbach 2011), citation-based measures for impact are still the subject of much scientometrics research. This includes new methods for evaluating research institutions (Prathap 2014), journals (Braun, Glänzel and Schubert 2006; Bergstrom 2007; Bergstrom, West and Wiseman 2008), and researchers (Bletsas and Sahalos 2009; Zhang 2009; Cormode *et al.* 2013; Bornmann and Marx 2013; Ferrara and Romero 2013; Klosik and Bornholdt 2014; Radicchi and Castellano 2013; Egghe 2014).

Apart from quantifying scholarly impact, other applications of citation analysis includes research user profiling (Kostoff *et al.* 2001), measuring diffusion characteristics of academic knowledge (Frandsen and Nicolaisen 2013; Liu and Rousseau 2014), analyzing social aspects of scientific research (Milard 2014), scientific information retrieval and indexing (Bradshaw 2003; Garfield 2006), analyzing history, structure and progress of scientific fields (Heneberg 2013; Shen *et al.* 2013; Velden and Lagoze 2013) and measuring interdisciplinarity of scientific fields (Rafols and Meyer 2009; Zitt and Cointet 2013).

Most of these methods reduce a citation to a single edge between the citing and cited paper and treat all the edges equally. This is clearly an oversimplification since all citations are not equal. As a simple extension, taking into account the number of times a paper is cited in the citing paper often does a better job of measuring the impact of the cited paper (Hou, Li and Niu 2011; Wan and Liu 2014a). Taking a further step in this direction requires acknowledging that research publications are linguistic artifacts, and often, the text around the reference anchor in a citing paper provides additional context.

This text around a citation anchor can be used to assess the attitude of the citing paper towards the cited paper. We refer to this as citation polarity. Aggregating the attitudes of all the citations to a paper can give us a quantitative measure of the attitude of the community towards that paper. Citation polarity is differentiated from citation purpose, which relates to the citer's motivation. Work in this area has mostly focused on understanding citation purpose, leading to various classification schemes for these (Small 1982; Swales 1990). Teufel, Siddharthan and Tidhar (2006) is a recent work in this vein. They annotate citation purpose and derive citation polarity labels from them. In this paper, we describe an annotated dataset we have developed that contains both citation purpose and citation polarity annotations for the same set of sentences. This allows us to analyze the relationship between these two annotations. We also show some use cases of how automatic citation polarity detection can be useful for scientometric applications.

Traditional text-based citation analysis is usually limited to citing text that explicitly talks about a cited paper and contains a reference anchor. However, often the discourse about a paper either continues beyond the explicit citing sentence or begins a few sentences before the citing sentence. These implicit citations, henceforth called *citation context*, can contain both evaluative and informative signals. In this paper, we describe a dataset we created for citation context detection, describe its properties, and demonstrate the use of citation context for summarization and

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Citing Sentence	Reference Scope	Citation Context	Polarity
Most works done to create English POS taggers henceforth, taggers, for example, include (Church 1988), (Kupiec 1992), (Brill 1992) and (Voutilainen et al 1992)	Most works done to create English POS taggers henceforth, taggers, for example, include (Church 1988).	Most works done to create English POS taggers henceforth, taggers, for example, include (Church 1988). The problem with this framework, however, is that such reliable corpora are hardly available due to a huge amount of the labor-intensive work required.	Negative
Recent work involves novel ways to employ annotated corpus in part of speech tagging (Church 1988, Derose 1988) and the application of mutual information statistics on the corpora to uncover lexical information (Church 1989).	Recent work involves novel ways to employ annotated corpus in part of speech tagging (Church 1988, Derose 1988).	Research on corpus-based natural language learning and processing is rapidly accelerating following the introduction of large on-line corpora, faster computers, and cheap storage devices. Recent work involves novel ways to employ annotated corpus in part of speech tagging (Church 1988, Derose 1988).	Positive

Fig. 1. Some citing sentences to a classic NLP paper (Church 1988), and how NLP-driven citation analysis can provide additional insights into the response of the community towards the paper.

sentiment detection. The complementary problem to citation context is the problem of reference scope resolution: a citing sentence might reference multiple papers and therefore, parts of a sentence may not be talking about a cited paper even if it contains a reference anchor to it. In this paper, we describe a dataset we created for this task and illustrate its usefulness for sentiment detection and scientific summarization.

Figure 1 shows two citing sentences to a classic NLP paper about part of speech tagging (Church 1988) and shows how each of these additional NLP components can help us extract signals from the citing text that references a specific target paper while minimizing the noise that inevitably results from using the text from scientific papers outside of their original context.

We use the ACL Anthology Network (AAN) (Radev *et al.* 2013) as a corpus for all of our experiments. AAN is suitable as a case study because it provides the full

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

text of papers published at most of the venues in the field of Natural Language Processing (NLP) and provides additional useful data such as a manually curated-citation network between all the papers in the corpus. In its current release, AAN contains more than 21,000 publications from 342 venues in NLP. It contains close to 122,000 citing sentences and provides an ideal test bed for experimentation with citation analysis methods.

We now discuss citation purpose and polarity, citation context extraction and reference scope extraction in detail. This is followed by a related work review and concluding remarks.

2 Citation purpose and polarity

The analysis of citation polarity and purpose can be extremely helpful in evaluating and summarizing the impact of papers. For our study, we first created an annotated dataset. We selected thirty papers from AAN that had different numbers of incoming citations and were consistently cited since they were published. These thirty papers received a total of about 3,500 citations from within AAN (average = 115 citation/paper, Min = 30, and Max =338). These citations come from 1,493 unique papers.

As described above, one of the goals of this project was to study the interplay of citation purpose and polarity. Therefore, we asked our annotators to mark each sentence for both purpose and polarity. The citation polarity labels were positive, negative, and neutral. These labels are defined in a slightly different way than their usual sense. A citation is marked positive if it either explicitly states a strength of the target paper or indicates that the work done in the target paper has been used either by the author or a third-party. It is also marked as positive if it is compared to another paper (possibly by the same authors) and deemed better in some way. A citation is marked negative if it explicitly points to a weakness of the target paper. It is also marked as negative if it is compared to another paper and deemed worse in some way (our criteria for marking negative citations is equivalent to Teufel *et al.* (2006)). A citation is marked as neutral if it is only descriptive.

For annotating citation purpose, we created a taxonomy consisting of six categories based on our study of similar taxonomies proposed in previous work (Spiegel-Rösing 1977; Teufel *et al.* 2006). We selected the categories that we believe are more important and useful from a bibliometric point of view, and the ones that can be detected through citation text analysis. We also tried to limit the number of categories by grouping similar categories proposed in previous work under one category. The six categories, their descriptions, and an example for each category are listed in Table 1.

We asked graduate students with good background in NLP to provide two annotations for each citation example. For the first annotation, we asked them to determine the purpose of citing the target reference by choosing from the six purpose categories that we described earlier. For the second annotation, we asked them to determine whether the citation is negative, positive, or neutral. As mentioned

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 1. *Annotation scheme for citation purpose, motivated by the work of Spiegel-Rösing (1977) and Teufel et al. (2006)*

Category	Description	Example
Criticizing	Criticism can be positive or negative. A citing sentence is classified as 'criticizing' when it mentions the weakness/strengths of the cited approach, negatively/positively criticizes the cited approach, negatively/positively evaluates the cited source.	Chiang (2005) introduced a constituent feature to reward phrases that match a syntactic tree but did not yield significant improvement.
Comparison	A citing sentence is classified as 'comparison' when it compares or contrasts the work in the cited paper to the author's work. It overlaps with the first category when the citing sentence says one approach is not as good as the other approach. In this case we use the first category.	Our approach permits an alternative to minimum error-rate training (MERT; Och, 2003);
Use	A citing sentence is classified as 'use' when the citing paper uses the method, idea or tool of the cited paper.	We perform the MERT training (Och, 2003) to tune the optimal feature weights on the development set.
Substantiating	A citing sentence is classified as 'substantiating' when the results, claims of the citing work substantiate, verify the cited paper and support each other.	It was found to produce automated scores, which strongly correlate with human judgements about translation fluency (Papineni *et al*., 2002).
Basis	A citing sentence is classified as 'basis' when the author uses the cited work as starting point or motivation and extends on the cited work.	Our model is derived from the hidden-markov model for word alignment (Vogel *et al*., 1996; Och and Ney, 2000).
Neutral	A citing sentence is classified as 'neutral' when it is a neutral description of the cited work or if it doesn't come under any of the above categories.	The solutions of these problems depend heavily on the quality of the word alignment (Och and Ney, 2000).

above, asking annotators to do both of these allows us to study the interrelationship between citation purpose and polarity.

The complete annotated dataset is available for download at http://clair.si.umich.edu/corpora/citation_sentiment_umich.tar.gz.

To estimate the inter-annotator agreement, we randomly picked 400 sentences and assigned them to an additional annotator who did not originally annotate the sentences. We used the Kappa coefficient (Cohen 1968) to measure the agreement.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 2. *Distribution of the citations belonging to different citation purpose categories across polarity categories*

Purpose Label	Neutral	Positive	Negative
Criticizing	0%	33%	67%
Comparison	67%	17%	15%
Use	26%	73%	0%
Substantiating	1%	99%	0%
Basis	20%	80%	0%
Neutral	98%	1%	0%

Since both purpose and polarity annotations are categorical, Kappa coefficient can be directly used. The agreements on the purpose and the polarity classification task were $K = 0.61$ and $K = 0.66$ respectively, which indicates *substantial agreement* on the (Landis and Koch 1977) scale. These additional annotations produced for inter-annotator agreement are not part of the dataset.

2.1 Data analysis

The distribution of the purpose categories in the data was: 14.7 per cent criticism, 8.5 per cent comparison, 17.7 per cent use, 7 per cent substantiation, 5 per cent basis, and 47 per cent other. The distribution of the polarity categories was: 30 per cent positive, 12 per cent negative, and 58 per cent neutral. The amount of negative citations we found in our data is on the higher side, but within the range reported by previous work; (White 2004) reports that the previous literature has found the proportion of negative citations to be between 1 per cent and 14 per cent, with the highest number being reported in Moravcsik and Murugesan (1975).

Next, we study the correlation between the counts of the different purpose and polarity categories as well as total number of citations. This is shown in Table 2. Here we see the distribution of each citation purpose category across different polarity categories. Some of these correlations are expected. Criticizing sentences are usually interpreted as negative in nature, substantiating sentences are interpreted as positive, and neutral in both purpose and polarity correlate well. It is interesting to see that comparison usually results in a neutral citation polarity, indicating the lack of negative results in the scientific community. Similarly, basis and use usually lead to a positive polarity, which agrees with previous work. Finally, citations that are labeled with polarity label negative tend to co-occur exclusively with the purpose labels criticizing and comparison, which provides an experimental validation of the heuristics previously used by Teufel *et al.* (2006) to derive polarity labels automatically from purpose labels.

We also computed the Pearson correlation coefficient between the counts of citations from the different categories that a paper received per year since its publication. We found that, on average, the correlation between positive and negative citations is negative (AVG P $= -0.194$) and that the correlation between the count of positive citations and the total number of citations is much higher than the

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

correlation between negative citations and total citations (AVG P = 0.531 for positive versus AVG P = 0.054 for negative).

Similarly, we noticed that there is a higher positive correlation between *Use* citations and total citations than in the case of both *Substantiation* and *Basis*. One possible explanation for this is that publications that present new algorithms, tools, or corpora that are used by the research community become popular with time and thus receive more and more citations. An experiment to confirm this would be to label the citing sentences for the most cited papers in our corpus with citation purpose and measure which purpose categories dominate the citing sentences to these papers. We will do this as part of future work.

2.2 Automatic classification

We now turn to the problem of automatically labeling citation purpose and polarity. We provide a summary of our methods and experiments below, more details can be found in Abu-Jbara, Ezra and Radev (2013).

2.2.1 Methodology

We use a supervised approach whereby a classification model is trained on a number of lexical and structural features extracted from a set of labeled citation contexts. Some of the features that we use to train the classifier are listed in Table 3.

We experimented with several classifiers including: Support Vector Machines (SVM), Logistic Regression (LR), and Naive Bayes. Our initial experimentation showed that SVM outperformed Logistic Regression and Naive Bayes. Therefore, we report the results of ten-fold cross validation using SVM classifier on our data. All the results have been tested for statistical significance using a two-tailed paired t-test.

For polarity classification, due to the high skewness in the data (more than half of the citations are neutral), we use two setups for binary classification. In the first setup, the citation is classified as *Polarized (subjective)* or *neutral (objective)*. In the second one, *subjective* citations are classified as *positive* or *negative*. We find that this method gives more intuitive results than using a three-way classifier.

2.2.2 Citation purpose classification evaluation results

Table 4 shows the precision, recall, and F1 for each of the six categories. It also shows the overall accuracy and the Macro-F measure.

The chi-squared evaluation of the features shows that both lexical and structural features are important. It also shows that among lexical features, the ones that are limited to the existence of a direct relation to the target reference (such as *closest* verb, adjective, adverb, subjective cue, etc.) are the most useful. This can be explained by the fact that restricting the features to having direct dependency relations introduces much less noise than other features (such as *Dependency Triplets*). Among the structural features, the count of references in the citation context was found to be most useful.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 3. *Some of the features used for citation purpose and polarity classification*

Feature	Description
Reference count	The number of references that appear in the citation context.
Is Separate	Whether the target reference appears within a group of references or separate (i.e. single reference).
Closest Verb / Adjective / Adverb	The lemmatized form of the closest verb/adjective/adverb to the target reference or its representative or any mention of it. Distance is measure based on the shortest path in the dependency tree.
Self Citation	Whether the citation from the source paper to the target reference is a self citation.
Contains 1st/3rd PP	Whether the citation context contains a first/third person pronoun.
Negation	Whether the citation context contains a negation cue. The list of negation cues is taken from the training data of the *SEM 2012 negation detection shared task (Morante and Blanco 2012).
Speculation	Whether the citation context contains a speculation cue. The list is taken from Quirk *et al.* (1985)
Closest Subjectivity Cue	The closest subjectivity cue to the target reference or its representative or any anaphoric mention of it. The list of cues is taken from OpinionFinder (Wilson *et al.* 2005)
Contrary Expressions	Whether the citation context contains a contrary expression. The list is taken from Biber (1988)
Section	The headline of the section in which the citation appears. We identify five title categorizes: (1) *Introduction, Motivation, etc.* (2) *Background, Prior Work, Previous Work, etc.* (3) *Experiments, Data, Results, Evaluation, etc.* (4) *Discussion, Conclusion, Future work, etc.*. (5) All other section headlines. Headlines are identified using regular expressions.
Dependency Relations	All the dependency relations that appear in the citation context. For example, $nsubj(outperform, algorithm)$ is one of the relations extracted from 'This algorithm outperforms the one proposed by. . . '. The arguments of the dependency relation are replaced by their lemmatized forms. These types of features have been shown to give good results in similar tasks (Athar and Teufel 2012a).

2.2.3 Citation polarity classification evaluation results

Table 5 shows the precision, recall, and F1 values for both the subjectivity classifier (that classifies citation into either subjective or objective) and the polarity classifier (that classifies citation into either positive or negative) along with the accuracy for each classifier. Overall numbers for each of the three categories (positive, negative, and neutral) can be found in Abu-Jbara *et al.* (2013).

The analysis of the features used to train this classifier using chi-squared analysis leads to the same conclusions about the relative importance of the features as described in the previous subsection. However, we noticed that features that are related to subjectivity (*Subjectivity Cues*, *Negation*, *Speculation*) are ranked higher which makes sense in the case of polarity classification.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 4. *Summary of Citation Purpose Classification Results (10-fold cross validation, SVM: Linear Kernel, c = 1.0). Substantiating is abbreviated as Subst*

	Criticism	Comparison	Use	Subst.	Basis	Other
Precision	53.0%	55.2%	60.0%	50.1%	47.3%	64.0%
Recall	77.4%	43.1%	73.0%	57.3%	39.1%	85.1%
F1	63.0%	48.4%	66.0%	53.5%	42.1%	73.1%
Accuracy: 70.5%						
Macro-F: 58.0%						

Table 5. *Summary of citation polarity classification results (10-fold cross validation, SVM: Linear Kernel, c = 1.0)*

	Subjectivity classifier		Polarity classifier	
	Neutral %	Subjective %	Positive %	Negative %
Precision	84.0	86.1	88.5	96.8
Recall	96.0	57.7	99.2	66.1
F1	89.6	69.1	93.5	78.5
	Accuracy: 84.44 %		Accuracy: 90.1 %	

2.3 Applications

2.3.1 Measuring research dynamics

Figure 2 shows the results of running our purpose classifier on all the citations to Papineni *et al.* (2002)'s paper about BLEU, an automatic metric for evaluating Machine Translation (MT) systems. The figure shows that this paper receives a high number of *Use* citations. This makes intuitive sense because this paper describes an evaluation metric that has been widely used in the MT area. This also demonstrates that the citation purpose profiles of research papers can be used to do a more graded citation analysis which allows us to measure not just the amount of impact a paper is having, but also the nature of its impact. The figure also shows that in the recent years, this metric has started to receive some criticizing citations that has resulted in a slight decrease in the number of use citations. Such a temporal analysis of citation purpose and polarity is useful for studying the dynamics of research. It can also be used to detect the emergence or de-emergence of research techniques.

Similarly, Figure 3 shows that Church (1988) received significant positive feedback during the 1990s and until early 2000s before it started to receive more negative feedback. This can be explained by the emergence of better statistical models for part-of-speech (POS) tagging (e.g. Conditional Random Fields, CRF; Lafferty, McCallum and Pereira 2001) that outperformed Church's approach. However, as indicated by the neutral citation curve, Church's work continued to be cited as a

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

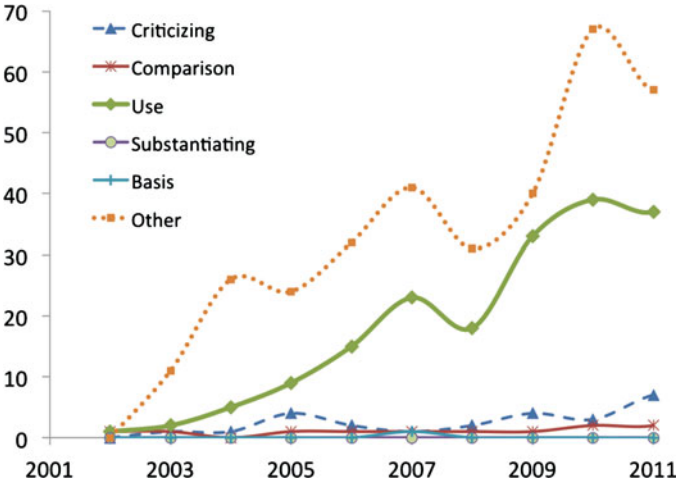


Fig. 2. (Colour online) Change in the purpose of the citations to Papineni *et al.* (2002). The values for the categories *Substantiating* and *Basis* for this paper are very close together and therefore, their lines are highly overlapping in the graph.

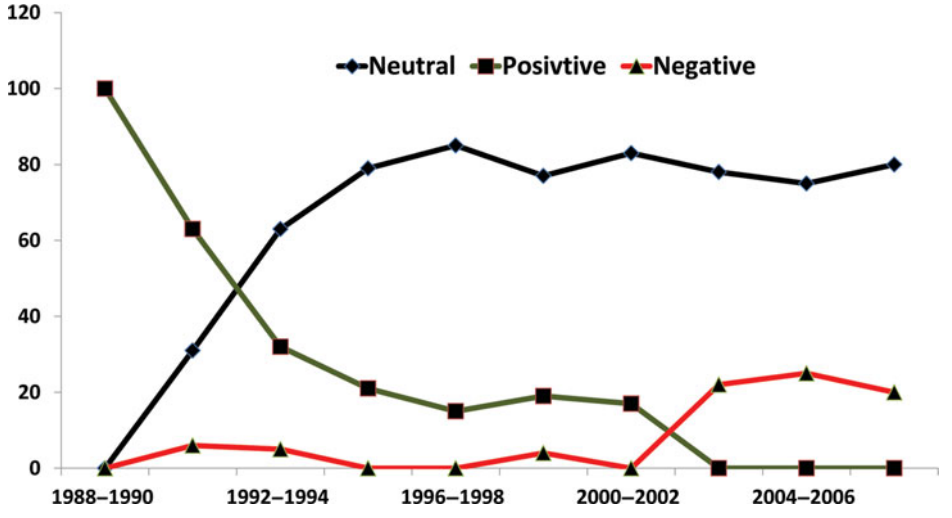


Fig. 3. (Colour online) Polarity for Church (1988).

classical pioneering research on the POS tagging task, but not as the state-of-the-art approach.

2.3.2 *Faceted summarization*

Citation-based summarization is based on the idea of using citing sentences to a target paper to generate a community-driven summary. Previous work has shown that such summaries can be more informative than the abstract of the target paper, because they take into account the response of the community to the target paper after the publication of the paper (Qazvinian and Radev 2008; Qazvinian, Radev and Özgür 2010). Existing citation-based summarization methods do not

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Positive citation summary sentences
Given a parse tree, we use a head percolation table (Magerman, 1995) to create the corresponding dependency structure.
For example, in the context of syntactic disambiguation, Black (1993) and Magerman (1995) proposed statistical parsing models based on decision-tree learning techniques, which incorporated not only syntactic but also lexical/semantic information in the decision-trees.
We use a statistical CFG parser to parse the English side of the training data, and extract dependency trees with Magerman's rules (1995).
Negative citation summary sentences
Comparison indicates that our best model is already better than the early lexicalized model of Magerman (1995).
Specifically, we construct an unlexicalized PCFG which outperforms the lexicalized PCFGs of Magerman (1995) and Collins (1996) though not more recent models, such as Charniak (1997) or Collins (1999).
Section 8.2 showed that the parsing models of Ratnaparkhi (1997), Jelinek et al. (1994), and Magerman (1995) can suffer from very similar problems to the "label bias" or "observation bias" problem observed in tagging models, as described in Lafferty, McCallum, and Pereira (2001) and Klein and Manning (2002).

Fig. 4. Sample sentences from faceted citation-based summary for the paper Magerman (1995).

take into account the attitude of citing sentences however, which can result in biased summaries. For example, any shortcomings of a generally positively cited paper are unlikely to appear in a summary, because the summarization algorithms depend on redundancy in the input to determine important facts.

The ability to assign polarities to citing sentences opens up new possibilities for more factored citation-based summarization. Citation-based summaries can be organized around the citation polarity or citation purpose, leading to *faceted summaries*. Figure 4 illustrates this idea by showing some sample sentences from a possible faceted summary for Magerman (1995), an early NLP paper about syntactic parsing. This summary was obtained by first classifying citing sentences to the paper into positive, negative, and neutral citations and then summarizing the positive and negative sentences separately. Given the set of input sentences for each category, each sentence was assigned a relevance score using Lexrank (Erkan and Radev 2004). Lexrank is a network-based content selection algorithm that works by first building a graph of all the document sentences in a cluster. The edges between corresponding nodes represent the cosine similarity between them. Once the network is built, the algorithm computes the salience of sentences in this graph based on their eigenvector centrality in the network. We hypothesize that such faceted summaries might provide a better overview of the strengths and weaknesses of a target paper compared to traditional citation-based summaries.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

The idea of faceted summaries for scholarly data links traditional scientific summarization with the highly active area of opinion summarization. However, more exploration is needed to create appropriate datasets and evaluation metrics for this task. This is because for evaluating faceted scientific summaries, we must come up with datasets and evaluation metrics that allow us to measure the coverage of facts in the summaries (achieved by traditional summarization evaluation metrics such as pyramid evaluation (Nenkova and Passonneau 2004)) as well as a balanced presentation of opinion (which is what datasets and metrics for opinion summarization tend to focus on Kim and Zhai (2009)). This would require creation of new datasets and evaluation metrics; we are currently pursuing research in this direction.

3 Citation context identification

Citing papers contain some explicit information about the paper being cited. The following example is an excerpt from a paper that contains information about Eisner's work on bottom-up parsers and the notion of span in parsing:

> Another use of bottom-up is due to **Eisner (1996)**, who introduced the notion of a span.

However, the citation to a paper may not always include explicit information about the cited paper:

> This approach is one of those described in **Eisner (1996)**

Although this sentence by itself does not provide any information about the cited paper, it suggests that its surrounding sentences describe the proposed approach in Eisner's paper:

> ... In an all pairs approach, every possible pair of two tokens in a sentence is considered and some score is assigned to the possibility of this pair having a (directed) dependency relation. Using that information as building blocks, the parser then searches for the best parse for the sentence. This approach is one of those described in **Eisner (1996)**.

We refer to such sentences that contain information about a specific secondary source but do not explicitly cite it as *context sentences*.

To build a corpus for studying such context sentences, we picked ten recently published papers from various areas in NLP and annotated them. Table 6 lists these papers together with their authors, publication year, number of references within AAN, and the number of sentences.

Each *annotation instance* in our setting corresponds to a paper-reference pair, and is a vector in which each dimension corresponds to a sentence and is marked with a *C* if it explicitly cites the reference, and with a one if it implicitly talks about it. All other sentences are marked with zeros. Table 7 shows a portion of two separate annotation instances of N03-1003 corresponding to two of its references. Our annotation has resulted in 203 annotation instances, each corresponding to one paper–reference pair. The complete annotation dataset is available for download at http://www-personal.umich.edu/ vahed/context-ext.html.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 6. *Papers chosen from AAN as source papers for the evaluation corpus, together with their authors, publication year, number of references (in AAN) and number of sentences. Papers marked with * are used to calculate annotation inter-judge agreement*

ACL-ID	Authors	Year	#Refs	#Sents
P08-2026	McClosky & Charniak	2008	8	102
N07-1025*	Mihalcea	2007	12	153
N07-3002	Wang	2007	14	74
P06-1101	Snow *et al.*	2006	9	138
P06-1116	Abdalla & Teufel	2006	10	231
W06-2933	Nivre *et al.*	2006	5	84
P05-1044	Smith & Eisner	2005	13	262
P05-1073	Toutanova *et al.*	2005	10	185
N03-1003	Barzilay & Lee	2003	13	203
N03-2016*	Kondrak *et al.*	2003	5	92

Table 7. *Part of the annotation for N03-1003 with respect to two of its references 'Lin and Pantel (2001)' (the first column) 'Shinyama et al. (2002)' (the second column). Cs indicate explicit citations, 1s indicate implicit citations and 0s are none*

L&P	S&al	Sentence
C	*C*	Jacquemin (1999) and Barzilay and McKeown (2001) identify phrase level paraphrases, while **Lin and Pantel (2001)** and **Shinyama *et al.* (2002)** acquire structural paraphrases encoded as templates.
1	1	These latter are the most closely related to the sentence-level paraphrases we desire, and so we focus in this section on template-induction approaches.
C	0	**Lin and Pantel (2001)** extract inference rules, which are related to paraphrases (for example, X wrote Y implies X is the author of Y), to improve question answering.
1	0	They assume that paths in dependency trees that take similar arguments (leaves) are close in meaning.
1	0	However, only two-argument templates are considered.
0	*C*	**Shinyama *et al.* (2002)** also use dependency-tree information to extract templates of a limited form (in their case, determined by the underlying information extraction application).
1	1	Like us (and unlike Lin and Pantel, who employ a single large corpus), they use articles written about the same event in different newspapers as data.
1	1	Our approach shares two characteristics with the two methods just described: pattern comparison by analysis of the patterns respective arguments, and use of nonparallel corpora as a data source.
0	0	However, extraction methods are not easily extended to generation methods.
1	1	One problem is that their templates often only match small fragments of a sentence.
1	1	While this is appropriate for other applications, deciding whether to use a given template to generate a paraphrase requires information about the surrounding context provided by the entire sentence.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 8. *Average κ coefficient as inter-judge agreement for annotations of two sets*

ACL-ID	Vector size	# Annotations	$\overline{\kappa}$
N07-1025	153	21	0.889 ± 0.30
N03-2016	92	8	0.853 ± 0.35

We also asked a neutral annotator to annotate two of our datasets that are marked with ∗ in Table 6. These additional annotations are used for measuring inter-annotator agreement and are not part of the dataset. For each paper–reference pair, the annotator was provided with a vector in which explicit citations were already marked with *C*s. The annotation guidelines instructed the annotator to look at each explicit citation sentence, read up to fifteen sentences before and after, and mark context sentences around that sentence with 1s. Next, the twenty-nine annotation instances done by the external annotator were compared with the corresponding annotations that we did, and the Kappa coefficient (κ) was calculated. To calculate κ, we ignored all explicit citations (since they were provided to the external annotator) and used the binary categories (i.e. one for context sentences, and zero otherwise) for all other sentences. Table 8 shows the annotation vector size (i.e. number of sentences), number of annotation instances (i.e. number of references), and average κ for each set. The average κ is above 0.85 in both cases, suggesting that the annotation process has a low degree of subjectivity and can be considered reliable.

3.1 Data analysis

First, we look at some statistics about the number of explicit citations a reference receives in a paper. Figure 5 shows a histogram corresponding to this distribution, which is calculated over all the references in the ten papers shown in Table 6.[1] It indicates that the majority of references get cited in only one sentence in a scientific article, while the maximum being nine in our collected dataset with only one instance (i.e. there is only one reference that gets cited nine times in a paper). Moreover, the data exhibits a highly positive-skewed distribution. This is illustrated on a log-log scale in Figure 6. This highly skewed distribution indicates that the majority of references get cited only once in a citing paper. The very small number of citing sentences cannot make a full inventory of the contributions of the cited paper, and therefore, extracting explicit citations alone without context sentences may result in information loss about the contributions of the cited paper.

Next, we investigate the distance between the context sentences and the closest citations. For each context sentence, we find its distance to the closest context sentence or explicit citation. Formally, we define the *gap* to be the number of sentences between a context sentence (marked with one) and the closest context

[1] For the histogram, we count the occurrence of all references whether they are in AAN or not, while the counts of references in Table 6 are given for references found in AAN only. This accounts for the difference in the total number of references in the earlier table and the histogram.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

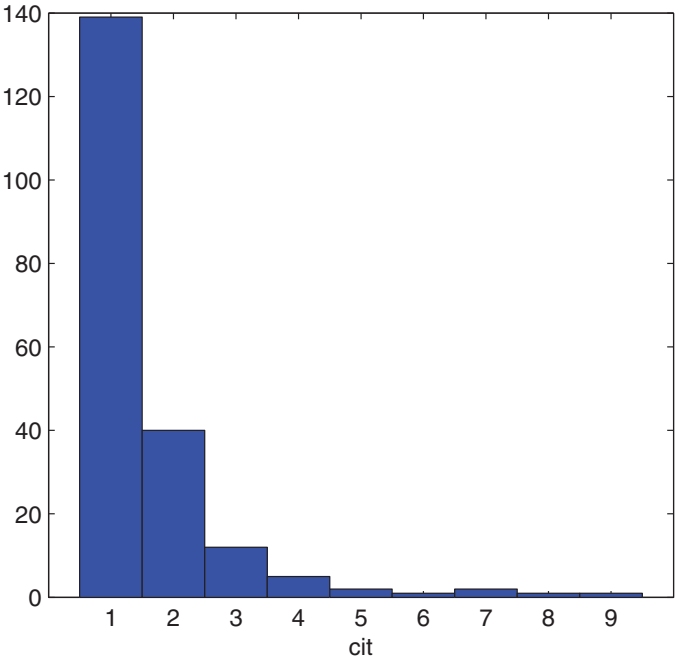


Fig. 5. (Colour online) Histogram of the number of different citations to all references (regardless of whether in AAN or not) found in the ten papers in our dataset.

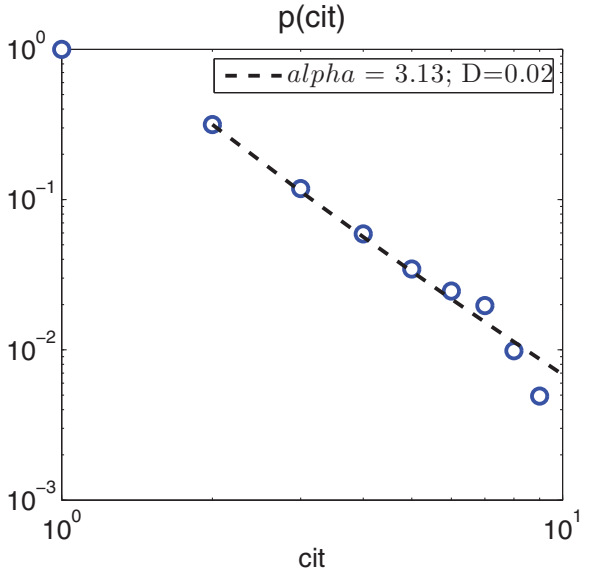


Fig. 6. (Colour online) The distribution observed for the number of different citations on a log-log scale.

sentence or explicit citation (marked with either C or one) to it. For example, the second column of Table 7 shows that there is a gap of size one in the ninth sentence in the set of context and citation sentences about both Shinyama *et al.* (2002) and

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 9. *The distribution of gaps in the annotated data*

Gap size	0	1	2	4	9	10	15	16
Instance	273	14	2	1	2	1	1	1

Lin and Pantel (2001). Table 9 shows the distribution of gap sizes in the annotated data. This distribution suggests that the majority of context sentences directly occur after or before a citation or another context sentence. However, it shows that gaps between sentences describing a cited paper actually exist, and a proposed method should have the capability to capture them.

3.2 Automatic classification

In this section, we sketch our methodology for automatically identifying the context sentences of a cited paper and summarize the experimental results. For more details, we refer the readers to Qazvinian and Radev (2010).

3.2.1 Methodology

The first step in building an automated system for identifying context sentences is creating informative features. We can use the structure of the problem to find such features. For example, since we are trying to find context sentences citing a specific target paper, sentences that have a high similarity to the target paper are likely to be citing sentences even if they donot have an explicit citation marker. Additionally, the fact that these sentences are all part of one discourse imposes certain constraints on their distribution in the text. The likelihood of a sentence being a context sentence depends upon how far it is from an explicit citing sentence for the target-cited paper: sentences closer to explicit sentences are more likely to be context sentences than those that are farther away. Thus, there are two classes of features that can help us determine context sentences: features inherent to a sentence itself, and features based on the relationship of a sentence with its neighboring sentences.

Markov Random Fields (MRFs) provide a natural way to model these two kinds of dependencies. An MRF is an undirected probabilistic graphical model where the predicted label of a node depends on both its intrinsic features as well as features based on its association with its neighbors. The intrinsic features that help us determine the label for each sentence can be modeled using the node potential function of the MRF, while the association between sentences can be modeled using the compatibility function.

We use three features in our node potential function. The first feature is based on whether the sentence is an explicit citing sentence. The second feature is based on whether the sentence matches two lexical patterns. The first pattern is a bigram pattern where the first term matches any of '*this; that; those; these; his; her; their; such; previous*' and the second term matches any of '*work; approach; system; method; technique; result; example*'. The second pattern matches any sentence that starts with

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

'*this; such*'. These patterns provide a heuristic way of identifying coreference between adjacent sentences. In future work, we hope to use standard coreference detection methods to model this more directly, as well as use other forms of discourse dependencies (for example, using explicit and implicit discourse markers from the Penn Discourse Treebank (Prasad *et al.* 2008)). The third feature is based on the lexical similarity of the sentence with the target cited paper, where the similarity is computed between the sentence and the full text of the cited paper.[2] The node potential function is then computed using an unweighted linear combination of these three features.

The dependencies between different sentences can be modeled in several ways depending on how we structure the dependency graph of the MRF. We can make a sentence depend on all the other sentences in the paper or we could make it depend on only a small window of sentences around it. Figure 7 shows the structure of the two MRFs at either extreme of the local dependency assumption. In Figure 7(a), each sentence depends on only its following and preceding sentences, while in Figure 7(b) each sentence depends on all the others. We refer to the former by $\mathbf{MRF_1}$ and to the latter by $\mathbf{MRF_n}$. Generally, we use $\mathbf{MRF_i}$ to denote a MRF in which each sentence is connected to i sentences before and after it. In the MRF framework, the dependency between different nodes is modeled using the compatibility function. We define this compatibility function in such a way that a sentence that is not labeled as a context sentence does not affect the labels of its neighbors in the graph. A sentence that is labeled as a context sentence affects its neighbors based on its lexical similarity with them. More details about the exact formulation of these MRF functions can be found in Qazvinian and Radev (2010). We experiment with three different structures of MRF: MRF_1, MRF_4, and MRF_n.

We compare our MRF-based method with three baselines. The first baseline, $\mathbf{B_{IR}}$, is an information retrieval-driven method. Given a paper-reference pair, for each explicit citation sentence, $\mathbf{B_{IR}}$ picks its preceding and following sentences if their similarities to that sentence are greater than a cutoff (the median of all such similarities), and repeats this for neighboring sentences of newly marked sentences. Intuitively, $\mathbf{B_{IR}}$ tries to find the best chain (window) around citing sentences.

As the second baseline, we use the hand-crafted discourse-based features used in MRF's potential function. This baseline, $\mathbf{B_{DISC}}$, marks as context those sentences that are within a particular distance (four in our experiments) of an explicit citation and match one of the discourse patterns mentioned earlier. After marking all such sentences, $\mathbf{B_{DISC}}$ also marks all sentences between them and the closest explicit citation no farther than four sentences away as context sentences.

Finally, as a third baseline, $\mathbf{B_{SVM}}$, we train a SVM classifier to label sentences as context/non-context. We use four features to train the **SVM** model. These four features comprise the three sentence level features used in MRF's potential function (similarity to reference, explicit citation, matching regular-expression) and a network level feature: distance to the closest explicit citation.

[2] The longer length of the full text of the cited paper compared to the citing sentence will cause the similarity scores to be generally lower, but since this is consistently done for all sentences, it leads to a fair comparison between different citing sentences.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

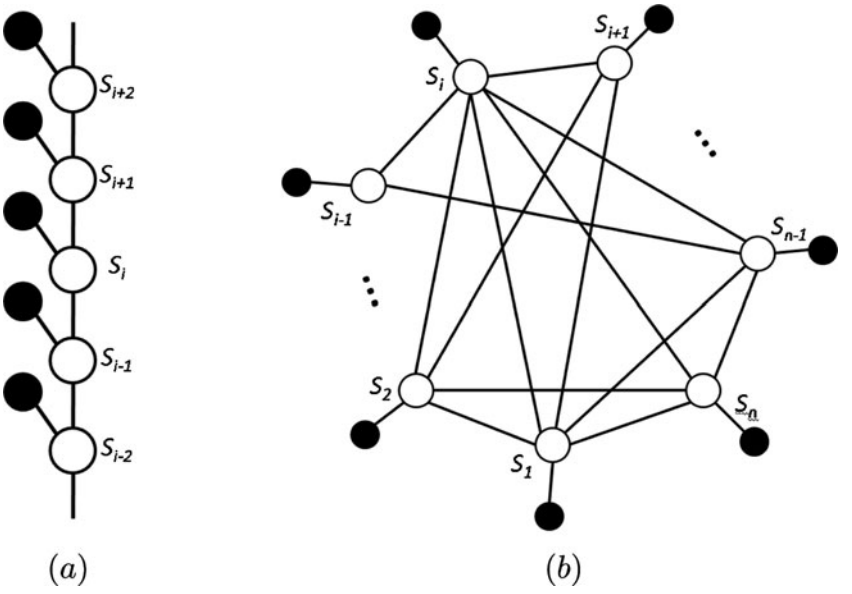


Fig. 7. The structure of the MRF constructed based on the dependency of non-adjacent sentences; (a) left, each sentence is independent on all other sentences given its immediate neighbors. (b) right, sentences have dependency relationship with each other regardless of their position.

3.2.2 *Evaluation results*

The evaluation of our methodology means to directly compare the output of our method with the gold standards obtained from the annotated data. Our methodology finds the sentences that cite a reference implicitly. Therefore, the output of the inference method is a vector, υ, of 1's, and 0's, whereby a 1 at element i means that sentence i in the source document is a context sentence about the reference while a 0 means an explicit citation or neither. The gold standard for each paper-reference pair, ω, is also a vector of the same format and dimensionality.

Precision, recall, and F_β for this task can be defined as

$$p = \frac{\upsilon \cdot \omega}{\upsilon \cdot \mathbf{1}}; \quad r = \frac{\upsilon \cdot \omega}{\omega \cdot \mathbf{1}}; \quad F_\beta = \frac{(1+\beta^2)p \cdot r}{\beta^2 p + r} \tag{1}$$

where $\mathbf{1}$ is a vector of 1's with the same dimensionality and β is a non-negative real number.

Table 10 shows the results of our experiments. We report F-score with β as three because higher values of β favor recall over precision. For this task, we are more interested in recall (e.g. if the citing sentence has five context sentences, we would want to get all of them, as opposed to getting only one and maximizing precision).

The best performing method is $\mathbf{MRF_4}$. Of the baseline methods, $\mathbf{B_{IR}}$ seems to do the best. This suggests that similarity to the explicit or implicit citing sentences is the strongest cue for finding context sentences. $\mathbf{B_{DISC}}$ does not do well on its own, but does improve the performance when combined with similarity-based features in the MRF-based classifiers.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 10. *Average $F_{\beta=3}$ for similarity-based baseline (**B_{IR}**), discourse-based baseline (**B_{DISC}**), a supervised method (**B_{SVM}**) and three MRF-based methods*

Paper	B_{IR}	B_{DISC}	B_{SVM}	MRF_1	MRF_4	MRF_n
P08-2026	0.441	0.237	0.249	0.470	**0.613**	0.285
N07-1025	0.388	0.102	0.124	0.313	**0.466**	0.138
N07-3002	0.521	0.339	0.232	**0.742**	0.627	0.315
P06-1101	0.125	0.388	0.127	0.649	**0.889**	0.193
P06-1116	0.283	0.104	0.100	0.307	**0.341**	0.130
W06-2933	0.313	0.100	0.176	0.338	**0.413**	0.160
P05-1044	0.225	0.100	0.060	0.172	**0.586**	0.094
P05-1073	0.144	0.100	0.144	0.433	**0.518**	0.171
N03-1003	0.245	0.249	0.126	**0.523**	0.466	0.125
N03-2016	0.100	0.181	0.224	0.439	**0.482**	0.185
Average	0.278	0.190	0.156	0.439	**0.540**	0.180

B_{IR} does much better than **B_{SVM}** as well. In addition to sentence level features, **B_{SVM}** has a feature that computes distance to the closest explicit citing sentence, but this feature is non-iterative and does not take into account lexical similarity. **B_{IR}**, on the other hand, works iteratively by adding additional sentences similar to the sentences already labeled as context sentences. It is difficult to create a feature that captures this iterative dependency in standard-supervised classifiers such as SVM. However, sequence-based classifiers such as MRFs are designed to model exactly this sort of dependency, and thus, it is not surprising that **MRF_1** and **MRF_4** do much better than all the other methods. These models are able to effectively capture sentence level features as well as the dependency between adjacent sentences.

Among the MRF-based methods, **MRF_4** does better than **MRF_1** which shows that sentences depend on more than one sentence on each side. However, the significantly worse performance of **MRF_n** suggests that dependencies on far away sentences do not improve results and in fact lead to a drop in performance, probably by leading to more false positives where sentences very far away from explicit citing sentences (and hence unlikely to be in their context) might be added as a context sentence due to an arbitrary similarity match.

3.3 Applications

3.3.1 Survey generation

Previous work in scientific summarization has shown the importance of citations in scientific domains and has indicated that citations include survey-worthy information (Teufel 2006; Elkiss *et al.* 2008; Qazvinian and Radev 2008; Mohammad *et al.* 2009). Here we show how context sentences add useful information to such summaries in addition to the information present in the explicit citations.

We use the data from Mohammad *et al.* (2009) that contains two sets of cited papers and corresponding citing sentences, one on question answering (QA) with ten papers and the other on dependency parsing (DP) with sixteen papers. For

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

... Naturally, our current work on question answering for the reading comprehension task is most related to those of (Hirschman et al., 1999; Charniak et al., 2000; Riloff and Thelen, 2000; Wang et al., 2000). **In fact, all of this body of work as well as ours are evaluated on the same set of test stories, and are developed (or trained) on the same development set of stories.** The work of (Hirschman et al., 1999) initiated this series of work, and it reported an accuracy of 36.3% on answering the questions in the test stories. **Subsequently, the work of (Riloffand Thelen, 2000) and (Charniak et al., 2000) improved the accuracy further to 39.7% and 41%, respectively. However, all of these three systems used handcrafted, deterministic rules and algorithms**...

...**The cross-model comparison showed that the performance ranking of these models was: U-SVM > PatternM > S-SVM > Retrieval-M.** Compared with retrieval-based [Yang et al. 2003], pattern-based [Ravichandran et al. 2002 and Soubbotin et al. 2002], and deep NLP-based [Moldovan et al. 2002, Hovy et al. 2001; and Pasca et al. 2001] answer selection, machine learning techniques are more effective in constructing QA components from scratch. **These techniques suffer, however, from the problem of requiring an adequate number of handtagged question-answer training pairs. It is too expensive and labor intensive to collect such training pairs for supervised machine learning techniques** ...

Fig. 8. A portion of the QA survey generated by LexRank using the context information. Bolded sentences are context sentences.

comparing different methods, we use pyramid evaluation (Nenkova and Passonneau 2004) using nuggets extracted manually from the citing sentences to these papers.

A nugget is defined as a specific fact that is found in the input sentences. These are found by using human generated regular expressions representing the facts that can then be matched directly to the text. Each nugget is assigned a weight based on the importance of the nugget, where the importance of a nugget is computed based on its frequency in the input text. Summaries that contain more highly weighted nuggets get better scores in the pyramid evaluation.

In pyramid evaluation, the nuggets are organized in a pyramid of order n. The top tier in this pyramid contains the highest weighted nuggets, the next tier contains the second highest weighted nuggets, and so on. The score assigned to a summary is the ratio of the sum of the weights of the nuggets it contains to the sum of weights of an optimal summary with the same number of nuggets. Pyramid evaluation allows us to capture how a method performs in terms of selecting sentences with the most highly weighted nuggets.

Both sets (QA and DP) contain nuggets extracted by experts from citation sentences. We use these nugget sets, which are provided in the form of regular expressions, to evaluate automatically generated summaries. To perform this experiment, we needed to build a new corpus that includes context sentences. For each citation sentence, we use our automatic context extraction method (described in Section 3.2.1) to extract the context sentences around it. Here, we limit the context

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 11. *Pyramid $F_{\beta=3}$ scores of automatic surveys of QA and DP data. The surveys are evaluated using nuggets drawn from citation texts (CT)*

	Citation survey	Context survey
Question answering	0.416	**0.634**
Dependency parsing	0.324	**0.379**

size to be four on each side. That is, we attach to a citing sentence any of its four preceding and following sentences that our method marks as context sentences. Therefore, we build a new corpus in which each explicit citation sentence is replaced with the same sentence attached to at most four sentence on each side. After building the context corpus, we use LexRank (Erkan and Radev 2004) to generate two QA and two DP surveys using the citation sentences only, and the new context corpus explained above.

We limit these surveys to be of a maximum length of 1,000 words. Figure 8 shows a portion of the survey generated from the QA context corpus. This example shows how context sentences add meaningful and survey-worthy information along with citation sentences. Table 11 shows the Pyramid $F_{\beta=3}$ score of automatic surveys of QA and DP data. Both surveys are evaluated using nuggets drawn from citation texts. In all evaluation instances, the surveys generated with the context corpora excel at covering nuggets.

We notice that adding context sentences tends to improve results for the topic of QA much more than DP. In general, the topic of DP seems to be more challenging for citation-based survey generation methods than the topic of QA. In earlier experiments (Qazvinian *et al.* 2013), we also found that the average pyramid score obtained by summaries generated by four human evaluators was 0.599 for QA while it was 0.413 for DP. Thus, a simple explanation for the difference in results between the two topics is that the factoid distribution in citing sentences (and hence context sentences) for QA is better compared to DP, leading to more incremental improvement as additional context information is added. For a better understanding of what topics might be more suited for such methods, a more detailed evaluation with a larger set of topics is needed. This will be part of our future work. Based on our understanding of the problem, this might have something to do with the structure of the lexical similarity networks generated for the different topics, which in turn is related to the distribution of different subtopics within these larger topics.

3.3.2 Polarity classification

To study the impact of using citation context in addition to the citing sentence on classification performance, we ran an additional polarity classification experiment. In the first polarity experiment (described in Section 2.2.3), we had only used the explicit citing sentences to extract the features that are used to train the classifiers.

For the second polarity experiment, we experimented with context information annotated by humans. For each of the citing sentences in our polarity experiments,

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 12. *Summary of Citation Polarity Classification Results with context sentences. (ten-fold cross validation, SVM: Linear Kernel, c = 1.0). Results under column wo/ctx are the results with only explicit citing sentences and no context sentences, while the results under column w/ctx are the results with context sentences included*

	Negative %		Positive %		Neutral %	
	wo/ctx	w/ctx	wo/ctx	w/ctx	wo/ctx	w/ctx
Precision	66.4	69.8	52.1	55.4	82.8	84.2
Recall	71.1	81.1	45.6	46.3	95.1	95.3
F1	68.7	75.0	48.6	50.4	88.5	89.4
Accuracy wo/ctx: 74.2 %, Accuracy w/ctx: 84.2 %						
Macro-F wo/ctx: 62.1 %, Macro-F w/ctx: 74.2 %						

we asked the human annotators to look at a window of four sentences around the explicit citation sentence and label the context sentences that are relevant to the target reference. We then extracted features from this gold context and ran a second round of classification.

Table 12 shows the results of both experiments. The results show that adding citation context improves performance, as shown in the increased F-Scores for each of the three categories (positive, negative, and neutral). Adding context improves precision by 1–3 points for each of the categories. The improvement in recall is unevenly distributed: for positive and neutral categories, recall improves a little, while for the negative category, we see an improvement of about ten points for recall. This supports the intuition about polarized citations that authors start their review of the cited work with an objective (neutral) sentence and then follow it with their criticism if they have any. Thus, taking these context sentences into account allows us to correctly label several negative sentences that were labeled as positive or neutral without this information.

4 Reference scope identification

When a reference appears in a scientific article, it is usually accompanied by a span of text that highlights the important contributions of the cited article. For example, sentence (1) below is a citing sentence that cites a paper by Philip Resnik and describes the problem Resnik addressed in his paper.

(1) **Resnik (1999)** addressed the issue of language identification for finding Web pages in the languages of interest.

The sentence above contains only one reference, so there is no ambiguity about what reference is being talked about. However, sentences that contain references to multiple papers are very common in scientific writing. For example, sentence (2) below contains three references.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

(2) Grefenstette and Nioche (2000) and Jones and Ghani (2000) use the web to generate corpora for languages where electronic resources are scarce, while **Resnik (1999)** describes a method for mining the web for bilingual texts.

In this sentence, the first fragment describes the contribution of Grefenstette and Nioche (2000) and Jones and Ghani (2000). The second fragment describes the contribution of Resnik (1999).

We use the term *Reference Scope* to refer to the fragments of a sentence that are relevant to a specific target paper. Two additional examples for reference scope are shown below, where sentences (4) and (5) are labeled for the target references Tetreault and Chodorow (2008), and Cutting *et al.* (1992) respectively. The underlined words are in the reference scope for the respective target papers.

(4) For example, **Tetreault and Chodorow (2008)** use a maximum entropy classifier to build a model of correct preposition usage, with 7 million instances in their training set, and Lee and Knutsson (2008) use memory-based learning, with 10 million sentences in their training set.

(5) There are many POS taggers developed using different techniques for many major languages such as transformation-based error-driven learning (Brill, 1995), decision trees (Black *et al.*, 1992), Markov model (**Cutting *et al.*, 1992**), maximum entropy methods (Ratnaparkhi, 1996) etc for English.

Our goal is to build an automatic classifier for reference scope detection. For this, we first built a dataset for reference scope from papers in AAN in the following way. In the complete set of citing sentences in AAN, 56 per cent contain two or more references and 44per cent contain one reference only. From this set, we randomly selected 3,500 citing sentences, each containing at least two references. The total number of references in this set of sentences is 19,591. We then asked graduate students with good background in NLP to provide three annotations for each sentence in the data set described above. First, we asked them to determine whether each of the references in the sentence was correctly tagged or not. Second, we asked them to determine, for each reference, whether it is a syntactic constituent in the sentence or not. Third, we asked them to determine and label the scope of the reference in each sentence which was marked as a target reference. We designed a user-friendly tool to collect the annotations from the students. The dataset is available for download at `http://clair.si.umich.edu/corpora/refscope_data.tar.gz`.

To estimate the inter-annotator agreement, we picked 500 random sentences from our data set and assigned them to an additional annotator who didn't originally annotate these sentences. We then measured inter-annotator agreement on the reference scope annotation task (not on the first two tasks of identifying correctly tagged references and identifying syntactic references). This is done by labeling each word as either one for inside or zero for outside the reference scope for each of the annotations and then computing the Kappa coefficient between the two annotations.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

The Kappa coefficient of agreement between the two sets of annotations on the scope identification task was $K = 0.61$. On Landis and Kochs scale (Landis and Koch 1977), this value indicates substantial agreement. These additional annotations are not part of the dataset.

4.1 Data analysis

Given a citing sentence and a target reference, we measured the number of words that are in the reference scope for the target reference. We then computed the proportion of words in the reference scope compared to the total number of words in the sentences. We found that on average, the reference scope for a given target reference contained only 57.63per cent of the original citing sentence (with standard deviation 24per cent). This tells us that citation analysis that does not take into account reference scope can be dealing with as much as 40per cent extraneous noise, or words that are not relevant to the target reference in question at all. This confirms our hypothesis that reference scope extraction should be an important part of citation analysis in order to measure the effects of citations properly.

Next, we look at the fragmentation patterns of reference scope in a citing sentences. Given a sentence with multiple references, do all the words relevant to a target reference appear together, or are they distributed throughout a sentence? Our analysis over the data shows that a reference scope for a given target is most likely to be concentrated over either one or two segments in the sentences. 45per cent of the citing sentences contain only a single connected reference scope for a target reference. 44per cent of the citing sentences contain two disconnected reference scopes and 10per cent contain three reference scopes. Higher number of disconnected reference scopes are rare and no sentences contain more than five reference scopes. This observation motivates one of our methodologies for finding reference scopes.

Finally, we measure the distance between fragments of reference scopes to a single target paper in terms of number of words. We found that this distance can have a very high variation. The average is 5.03 words with a standard deviation of 4.87.

4.2 Automatic classification

We now provide a sketch of the methods we explored for automatically identifying the scope of a given reference within a citing sentence. More details about the methodology and pre-processing steps are available in Abu Jbara and Radev (2012).

4.2.1 Methodology

Our general problem formulation is to label each word in the citing sentence as either belonging to the reference scope of a specific target reference or not. One of the most important clues about whether a word belongs to the reference scope is provided by the positional and syntactic relationship of the word with the target reference anchor. One approach would be to use a set of features for each word and learn a supervised classifier that can label each word as ‘inside’ or ‘outside’ the

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 13. *The features used for word classification and sequence labeling*

Feature	Description
Distance	The distance (in words) between the word and the target reference.
Position	This feature takes the value one if the word comes before the target reference, and zero otherwise.
Segment	After splitting the sentence into segments by punctuation and coordination conjunctions, this feature takes the value one if the word occurs in the same segment with the target reference, and zero otherwise.
Part of speech tag	The part of speech tag of the word, the word before, and the word after.
Dependency Distance	Length of the shortest dependency path (in the dependency parse tree) that connects the word to the target reference or its representative. It has been shown in previous work on relation extraction that the shortest path between any two entities captures the information required to assert a relationship between them (Bunescu and Mooney 2005)
Dependency Relations	This item includes a set of features. Each features corresponds to a dependency relation type. If the relation appears in the dependency path that connects the word to the target reference or its representative, its corresponding feature takes the value one, and zero otherwise.
Common Ancestor Node	The type of the node in the syntactic parse tree that is the least common ancestor of the word and the target reference.
Syntactic Distance	The number of edges in the shortest path that connects the word and the target reference in the syntactic parse tree.

reference scope based on these features. The set of features that we used to train our classifier are listed in Table 13. This is our simplest approach, which we call the *Word Classification* approach.

The *Word Classification* approach does not take into account the fact that the label of a word can depend on the labels of words around it. These relationships can be modeled using a sequence classifier; we use CRF to explore this. In CRF, the predicted label for a word depends on both the features for the word (which are the same as used in *Word Classification*) as well as the words around it. This second approach is our *Sequence Labeling* approach.

Both of the above mentioned approaches ignore the fact that words in a sentence are organized in a hierarchical syntactic structure. Thus, it might be useful to model the classification problem at units of higher granularity than words. One way to do this is to use a chunking tool to identify noun groups, verb groups, preposition groups, adjective groups, and adverb groups in the sentence. Each such group (or chunk) forms a segment. If a word does not belong to any chunk, it forms a singleton segment by itself. Labels are assigned by first using the *Sequence Labeling* approach to label the individual words in the sentence and then assigning to each segment the majority label of the words it contains. This third approach is our *Segment Classification* approach.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Table 14. *Results of scope identification using the different algorithms described in the paper*

Method	Accuracy	Precision	Recall	F1-score
AR-2011	54.0%	63.3%	33.1%	41.5%
Word classification	74.9%	74.5%	93.4%	82.9%
Sequence labeling	78.2%	80.1%	**94.2**%	86.6%
Segment classification	**81.8%**	**81.2%**	93.8%	**87.0%**

Finally, we also compare our systems with the baseline approach from Abu-Jbara and Radev (2011). In this approach, we first parse the citing sentence using a link grammar parser and extract as reference scope the smallest subtree rooted at a sentence node that contains the target reference anchor. If the text corresponding to this subtree is not grammatical, the second smallest subtree rooted at a sentence node is tried and so on. We call this approach AR-2011.

4.2.2 Evaluation results

Table 14 reports the precision, recall, F1-score, and accuracy for all the methods in ten-fold cross validation. All the metrics were computed at the word level. All the methods outperform the baseline method AR-2011. We also notice that the CRF-based *sequence labeling* method performs significantly better than the *word classification* method, corroborating our intuition that the labels of neighboring words are dependent.

The *segment classification* method outperforms both *word classification* and *sequence labeling* methods, showing that syntactic structure provides useful information for the task. On the other hand, the baseline method AR-2011 that relies exclusively on syntax-driven heuristics performs worse than all other methods. Thus, the combination of a feature-driven model along with syntax-based heuristics provides the best approach for this task. This suggests that for tasks of this nature, while machine learning methods perform well, good hybrid algorithms that also employ appropriate syntactic heuristics derived from an understanding of the structure of the problem can improve results further.

4.3 Applications

4.3.1 Focused Summarization

Reference scope extraction provides another set of possibilities for improving citation summarization by building more focused summaries complementary to the improvement provided by faceted summaries which are more comprehensive but can still contain irrelevant information due to the presence of multi-reference sentences.

The simplest way in which reference scope extraction can be used to improve citation summaries is by using it as a post-processing step. Thus, while the input

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

sentences for the salience model (e.g. Lexrank) are still full citation sentences as in previous work, once the model produces the set of most salient sentences, they can be run through a reference scope extraction module that can preserve only the sentence segments that are relevant to the target paper being summarized. Figure 9 illustrates how this might benefit summarization. The left column shows the first two sentences from the citation summaries generated for two different papers: Och and Ney (2003) which introduced Giza++ and Yarowsky (1995) which introduced bootstrapping for word sense disambiguation. The right column in the table shows the same sentences after applying reference scope extraction.

It is easy to see how reference scope-based post-processing can help provide more focus to the summaries. For the summary in Och and Ney (2003), it removes potentially irrelevant details about grow-diag-final-and method from the first sentence and about Moses and SRILM from the second sentence. Similarly, for Yarowsky (1995), it removes fragments about supervised sense disambiguation from the first sentence and details of other tasks using bootstrapping algorithms from the second sentence. It is possible that the user might not want to completely remove these details from sentences, in which case reference scopes can be highlighted while keeping the full citation sentence to emphasize the text fragments relevant to the target paper being summarized.

A second possibility is to use reference scope extraction as a pre-processing step for summarization. In this method, reference scopes are extracted from input citing sentences and the salience model is applied on them. This might lead to better summaries because only the text fragments relevant to the target paper are used by the salience model to assign importance to sentences, leading to an improved signal versus noise ratio. The final summary can consist of only the most salient reference scopes or the full sentences that consist of the most salient reference scopes as scored by the salience model.

Our future work includes creating a dataset of reference scope annotations geared towards summarization so that we can experiment with each of these possibilities. The evaluation strategy for this would be a mixture of automatic evaluation to measure the information content of reference scope-based summaries and human evaluation to measure the readability of such summaries.

4.3.2 Citation sentiment analysis

Reference scope extraction can also help us do more accurate citation-based sentiment analysis. Sentence (1) below shows an examples of this case.

(1) Cohn and Lapata (2008) used the GHKM extraction method (Galley *et al.* 2004), which is limited to constituent phrases and thus produces a reasonably small set of syntactic rules.

If the target reference is Cohn and Lapata (2008), only the underlined segment should be used for feature extraction. The limitation stated in the second segment of sentence is referring to Galley *et al.*, (2004). Similarly, consider citing sentence (2):

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Citation summary	Reference scoped summary
(Och and Ney 2003)	
The combined training corpus from which we extracted our grammar consisted of 123,609 sentence pairs, which was then filtered for length and aligned using the GIZA++ implementation of IBM Model 4 (Och and Ney, 2003) to obtain one-to-many alignments in either direction and symmetrized using the grow-diag-final-and method (Koehn et al., 2003).	The combined training corpus from which we extracted our grammar consisted of 123,609 sentence pairs, which was then filtered for length and aligned using the GIZA++ implementation of IBM Model 4 <> to obtain one-to-many alignments in either direction
The translation system is a factored phrase-based translation system that uses the Moses toolkit (Koehn et al., 2007) for decoding and training, GIZA++ for word alignment (Och and Ney, 2003), and SRILM (Stolcke, 2002) for language models.	The translation system is a factored phrase-based translation system that uses <> GIZA++ for word alignment
(Yarowsky 1995)	
Many corpus based methods have been proposed to deal with the sense disambiguation problem when given definition for each possible sense of a target word or a tagged corpus with the instances of each possible sense, e.g., supervised sense disambiguation (Leacock et al. , 1998), and semi-supervised sense disambiguation (Yarowsky, 1995).	Many corpus based methods have been proposed to deal with the sense disambiguation problem when given definition for each possible sense of a target word or a tagged corpus with the instances of each possible sense, e.g., <> semi-supervised sense disambiguation
Many bootstrapping algorithms have been proposed for a variety of tasks: word sense disambiguation (Yarowsky, 1995; Abney, 2004;), information extraction (Hearst, 1992; Riloff and Jones, 1999; Thelen and Riloff, 2002; Pantel and Pennacchiotti, 2006;), named entity recognition (Collins and Singer, 1999), part-of-speech tagging (Clark et al., 2003), and statistical parsing (Steedman et al., 2003; McClosky et al., 2006;).	Many bootstrapping algorithms have been proposed for a variety of tasks: word sense disambiguation

Fig. 9. Top sentences from the summaries obtained for two papers along with their reference scopes. The token <> is used to separate non-adjacent fragments of the same reference scope.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

(2) Automatic text summarization approaches have offered reasonably well-performing approximations for identifying important sentences (Lin and Hovy, 2002; Schiffman *et al*., 2002; Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Daume III and Marcu, 2006) but, not surprisingly, text (re)generation has been a major challenge despite some work on sub-sentential modification (Jing and McKeown, 2000; Knight and Marcu, 2000; Barzilay and McKeown, 2005).

Clearly the sentence expresses two different sentiments towards the first and second group of papers. A sentiment extraction system that uses this sentence for analyzing the impact of a specific paper, say Lin and Hovy (2002), may additionally aggregate attitudes and information about a completely different set of papers. Reference scope extraction can help us use only the segments of a citation sentence relevant to a specific target to determine the citation polarity and purpose for the sentence. This would lead to more accurate assessment of the attitude of the community towards the target paper. In future work, we plan to measure the effect of using reference scopes on citation sentiment analysis.

5 Related work

Our work is related to a large body of research on citations. Studying citation patterns and referencing practices has interested researchers for many years (Hodges 1972; Garfield Sher and Torpie 1984). White (2004) provides a good survey of the different research directions that study or use citations.

Several research efforts have focused on studying the different purposes for citing a paper (Garfield 1964; Weinstock 1971; Chubin and Moitra 1975; Moravcsik and Murugesan 1975; Bonzi 1982). Bonzi (1982) studied the characteristics of citing and cited works that may aid in determining the relatedness between them. Garfield (1964) enumerated several reasons why authors cite other publications, including alerting researchers to forthcoming work, paying homage to the leading scholars in the area, and citations which provide pointers to background readings. Weinstock (1971) adopted the same scheme that Garfield proposed in her study of citations.

Spiegel-Rösing (1977) proposed thirteen categories for citation purpose based on her analysis of the first four volumes of Science Studies. Some of them are: cited source is the specific point of departure for the research question investigated, cited source contains the concepts, definitions, interpretations used, and cited source contains the data used by the citing paper. Nanba and Okumura (1999) came up with a simple schema composed of only three categories: *Basis*, *Comparison*, and other *Other*. They proposed a rule-based method that uses a set of statistically selected cue words to determine the category of a citation. They achieved an F-score of 83per cent on a dataset of fifty citing sentences for these types. The high scores compared to our results on citation purpose classification can be attributed to a simpler task and limited data. It must be noted that their main focus was scientific paper summarization, which justifies the simpler schema. Teufel *et al.* (2006) did a much larger evaluation on 548 citation sentence in their work on citation function classification. They adopted twelve categories from Spiegel–Rosing's taxonomy and

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

trained an SVM classifier to label each citing sentence with exactly one category. Further, they mapped the twelve categories to four top level categories namely: weakness, contrast (four categories), positive (six categories), and neutral. Their performance in terms of F-score is similar to ours (between 28per cent to 86per cent).

The polarity (or sentiment) of a citation has also been studied previously. Previous work showed that positive and negative citations are common, although negative citations might be expressed indirectly or in an implicit way (Ziman 1968; MacRoberts and MacRoberts 1984; Thompson and Yiyun 1991). Athar (2011) addressed the problem of identifying sentiment in citing sentences. They used a set of structure-based features to train a machine learning classifier using annotated data with 8,736 manually annotated citations. Their SVM-based approach got a macro F-score of 76.4 per cent, which is higher than the macro F-score achieved by our system (62.1 per cent). This difference can be attributed to their use of effective sentence structure-based features.

Citation purpose and relevance has been used for doing scientometric analysis in a number of different fields. Li *et al.* (2014) use citation motivation to study science linkage: a widely used patent bibliometric indicator to measure patent linkage to scientific research based on the frequency of citations to scientific papers within the patent. Liu *et al.* (2014a) use citation relevance-based main path analysis for tracing main paths of legal opinions and show that relevancy information helps main path analysis uncover legal cases of higher importance. Cheang *et al.* (2014) use citation classification to do a multidimensional evaluation of thirty-nine selected management journals.

Other uses of citation purpose include a study by Bonzi and Snyder (1991) to understand citation purpose in the context of self-citation in natural sciences. Wan and Liu (2014b) present a regression method for automatically estimating the strength value of each citation and show the estimated values can achieve good correlation with human-labeled values.

We now look at some work on linguistic analysis of citation text. Nakov, Schwartz and Hearst (2004) proposed the use of citation text as a tool for semantic interpretation of bioscience text and propose several applications. Ding *et al.* (2014) introduced the notion of Citation Content Analysis (CCA) and discussed the nature and purposes of CCA along with potential procedures to conduct CCA. Halevi and Moed (2013) present a citation context analysis for the journal of infometrics. Zhao and Strotmann (2014) also analyze the feasibility, benefits, and limitations of in-text author citation analysis and test how well it works compared with traditional author citation analysis using citation databases. Angrosh, Cranefield and Stanger (2013) present a dataset for citation context sentences and present a model for citation context identification based on Conditional Random Fields.

One of the important uses of citation context is for scientific summarization. Nanba and Okumura (1999) use the term *citing area* to refer to the same concept as citation context. In Nanba, Kando and Okumura (2004), they use their algorithm to improve citation type classification and automatic survey generation. Kaplan, Iida and Tokunaga (2009) present a method for identifying citation contexts based on coreference analysis and their use for research paper summarization. They report

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

F-scores of about 69 per cent, which is higher than the numbers we obtained in our experiments (54 per cent). This can be attributed to their better modeling of coreference chains, which are modeled using a simple heuristics pattern-based matching method in our system. These two approaches are complementary and it is likely that better results can be obtained by combining our MRF-based approach with features based on coreference chains.

Citation context has also been used for the literature retrieval models for scientific domains. Liu *et al.* (2014b) designed a retrieval system for the PubMed Central database using citation contexts. Yin, Huang and Li (2011) similarly used citation context for the task of literature retrieval in the biomedical domain.

Athar and Teufel (2012b) observed that taking the context into consideration when judging sentiment in citations increases the number of negative citations by a factor of three. They also proposed two methods for utilizing the context. Their experiments surprisingly gave negative results and showed that classifying sentiment *without* considering the context achieves better results. They attributed this to the small size of their training data and to the noise that including the context text introduces to the data. However, in our experiments, we have seen that adding context leads to a large improvement in macro F-score (close to 12 per cent). This shows that even though doing context detection jointly with sentiment detection does not lead to better results (as found by them), context does provide useful information for sentiment detection (as found by us), and it is worthwhile to pursue this direction in future work. This is further supported by encouraging results presented in Athar and Teufel (2012a), where the authors present a method for automatically identifying all the mentions of the cited paper in the citing paper. They show that considering all the mentions improves the performance of detecting sentiment in citations.

6 Conclusions and future work

Using citations and citing sentences for analyzing and understanding the impact of research publications is an illustration of the idea of collective intelligence (Surowiecki 2004): how collecting independent, diverse perspectives from several different sources and aggregating them can lead to a much better picture than any single one of them can provide. The citations do provide independent qualitative and informative assessments of research papers, but the challenging aspect of using citing sentences for quantifying such assessments lies in creating reliable aggregation methods. This is a challenging issue because it requires us to use the citations and citing sentences out of their original context: the primary goal of citations for a citing author is to support the argumentation for their own work. The discourse properties of the language used for citing other papers derive from this motivation, and therefore, using them to compute aggregate assessments of the cited research works requires methods for separating the signal from the noise. In this paper, we argue that NLP-driven methods allow us to build innovative methods to achieve this goal.

Towards this end, we have focused on three specific tasks in this paper. We discuss the use of citation purpose and polarity for computing aggregate statistics

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

that represent the attitude of the scientific community towards a specific target paper. We then describe two important linguistic issues that come up when doing this kind of aggregation. First, there might be implicit sentences in the citing paper that talk about the target paper but may not contain an explicit reference to the target paper. This is the problem of citation context detection. Second, a single citing sentence might contain references to multiple papers apart from the target paper, and only a small segment of the sentence might be relevant to the target paper. This is the problem of reference scope detection. We present annotated datasets that we have created to study both these phenomenon, show how these annotations can be helpful in doing more focused analyses of target papers, and describe methodologies for automatic classification of both citation context and reference scope.

Most past research by other groups as well as our own work till now has focused on tackling these problems independently. Thus, separate datasets have been created for each of these tasks and various methods were explored. This has been useful in order to understand the structure of these problems and getting useful performance. In this paper, we have argued that a pipeline that labels sentences in scientific papers with each of these annotations (citation purpose and polarity, citation context, and reference scope) can be used to support several applications including scientific summarization and measuring research dynamics. Additionally, there is some interaction between these tasks as well. For example, citation context detection seems to improve citation purpose and polarity classification. Thus, future research should be geared towards building integrated datasets and pipelines that can help us create and evaluate these different tasks together. We now outline some more concrete steps that would help us get closer to this goal.

Datasets: One of the first tasks is to create a comprehensive and integrated dataset for topics in NLP and possibly other areas. Some other areas that have publicly available bibliographic datasets that can be used are Pubmed (Biomedical), Arxiv (High Energy Physics), and DBLP (Computer Science). For each of these areas, a comprehensive set of topics and a representative set of papers in each topic should be chosen. The citing sentences for each of the papers in this dataset should then be annotated with citation purpose and polarity, citation context, and reference scope. This dataset, in addition to helping us evaluate these three tasks consistently, would also serve as a starting point for other data annotation for evaluating applications such as summarization.

Integrated methodology: We have seen that sequence labeling methods seem to work well for both citation context identification and reference scope extraction. Both of these tasks can be considered aspects of a more general task: given a paper A that cites paper B, find all the text segments in A that directly relate to B (where the text segments can be sentences or sub-sentential units such as phrases or even words). Once a suitable classification methodology for this is found, it should be then possible to label each of the text segments with purpose and polarity and then recursively build up the sentiment of higher level linguistic units, a method for sentiment classification that has been shown to be effective in recent NLP work (Socher *et al.* 2013).

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

System deployment: Laboratory evaluation as presented in this paper is a good starting point for tackling these tasks, but the ultimate value of these methods can only be judged in real world systems that researchers can use. Portals such as ACL AAN[3] already provide some basic annotations to users such as citation profiles and citation summaries. An important part of future work should be to deploy the applications of these tasks such as citation polarity and purpose profiles, faceted summaries, and summaries based on citation context and reference scope on these portals and collect user feedback. This would help align the tasks in order to better serve the needs of researchers, which is the ultimate goal of this work.

References

Abu-Jbara, A., Ezra, J., and Radev, D. R. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *HLT-NAACL*, Atlanta, Georgia, USA, Association for Computational Linguistics, pp. 596–606.

Abu-Jbara, A., and Radev, D. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, Association for Computational Linguistics, pp. 500–9.

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*, Montréal, Canada, Association for Computational Linguistics, pp. 80–90.

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 - Volume 144*, AWC '13, Darlinghurst, Australia, Australia, Australian Computer Society, Inc, pp. 59–68.

Athar, A. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, Portland, OR, USA, Association for Computational Linguistics, pp. 81–7.

Athar, A., and Teufel, S. 2012a. Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, Jeju Island, Korea, Association for Computational Linguistics, pp. 18–26.

Athar, A., and Teufel, S. 2012b. 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*, NAACL HLT '12, Montréal, Canada, Association for Computational Linguistics, pp. 597–601.

Bergstrom, C. E. 2007. Measuring the value and prestige of scholarly journals. *Coll Res Libr News* **68**(5): 314–316.

Bergstrom, C. T., West, J. D., and Wiseman, M. A. 2008. The EigenfactorTM metrics. *Journal of Neuroscience* **28**(45): 11433–4.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge, Cambridge University Press.

Bletsas, A., and Sahalos, J. N. 2009. Hirsch index rankings require scaling and higher moment. *Journal of the American Society for Information Science and Technology* **60**(12): 2577–86.

Bonzi, S. 1982. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science* **33**(4): 208–16.

[3] http://clair.eecs.umich.edu/aan/

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Bonzi, S., and Snyder, H. W. 1991. Motivations for citation: a comparison of self citation and citation to others. *Scientometrics* **21**(2): 245–54.

Bornmann, L., and Marx, W. 2013. Standards for the application of bibliometrics in the evaluation of individual researchers working in the natural sciences. *ArXiv e-prints*.

Bornmann, L., and Marx, W. 2014. The wisdom of citing scientists. *Journal of the Association for Information Science and Technology* **65**(6): 1288–92.

Bradshaw, S. 2003. Reference directed indexing: redeeming relevance for subject search in citation indexes. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries,* Trondheim, Norway.

Braun, T., Bujdosó, E., and Schubert, A. 1987. Literature of Analytical Chemistry: A Scientometric Evaluation. *Boca Raton, FL: CRC Press*.

Braun, T., Glänzel, W., and Schubert, A. 2006. A hirsch-type index for journals. *Scientometrics* **69**(1): 169–173.

Brody, T., Harnad, S., and Carr, L. 2006. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology* **57**(8): 1060–72.

Bunescu, R., and Mooney, R. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, Association for Computational Linguistics, pp. 724–31.

Cheang, B., Chu, S. K. W., Li, C., and Lim, A. 2014. A multidimensional approach to evaluating management journals: refining pagerank via the differentiation of citation types and identifying the roles that management journals play. *Journal of the Association for Information Science and Technology* **65**(12): 2581–91.

Chubin, D. E., and Moitra, S. D. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science* **5**(4): 423–41.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136–43, Austin, Texas, USA. Association for Computational Linguistics.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**(4): 213–20.

Cormode, G., Ma, Q., Muthukrishnan, S., and Thompson, B. 2013. Socializing the h-index. *Journal of Informetrics* **7**(3): 718–21.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., and Zhai, C. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology* **65**(9): 1820–33.

Egghe, L. 2014. A good normalized impact and concentration measure. *Journal of the Association for Information Science and Technology* **65**(10): 2152–54.

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? *J. Am. Soc. Inf. Sci. Technol.* **59**(1): 51–62.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22:457–79.

Eysenbach, G. 2011. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4).

Ferrara, E., and Romero, A. E. 2013. Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the h-index. *Journal of the American Society for Information Science and Technology* **64**(11): 2332–39.

Frandsen, T. F., and Nicolaisen, J. 2013. The ripple effect: citation chain reactions of a nobel prize. *Journal of the American Society for Information Science and Technology* **64**(3): 437–47.

Garfield, E. 1964. Can citation indexing be automated? *Statistical Assoc. Methods for Mechanized Documentation, Symposium Proceedings*. Washington, US.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Garfield, E. 2006. Citation indexes for science. a new dimension in documentation through association of ideas. *International Journal of Epidemiology* **35**(5):1123–27.

Garfield, E., Sher, I. H., and Torpie, R. J. 1984. *The Use of Citation Data in Writing the History of Science*. Institute for Scientific Information Inc., Philadelphia, Pennsylvania, USA.

Gorraiz, J., Gumpenberger, C., and Schlögl, C. 2014. Usage versus citation behaviours in four subject areas. *Scientometrics* **101**(2): 1077–95.

Halevi, G., and Moed, H. F. 2013. The thematic and conceptual flow of disciplinary research: A citation context analysis of the journal of informetrics , 2007. *Journal of the American Society for Information Science and Technology* **64**(9): 1903–1913.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., and Larivière, V. 2014. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology* **65**(4): 656–69.

Heneberg, P. 2013. Lifting the fog of scientometric research artifacts: On the scientometric analysis of environmental tobacco smoke research. *Journal of the American Society for Information Science and Technology* **64**(2): 334–44.

Hodges, T. L. 1972. *Citation Indexing-its Theory and Application in Science, Technology, and Humanities*. Ph.D. Thesis, University of California at Berkeley.

Hou, W.-R., Li, M., and Niu, D.-K. 2011. Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution: citation frequency of individual articles in other papers more fairly measures their scientific contribution than mere presence in refere. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* **33**(10): 724–7.

Jonkers, K., Derrick, G. E., Lopez-Illescas, C., and Van den Besselaar, P. 2014. Measuring the scientific impact of e-research infrastructures: a citation based approach? *Scientometrics* **101**(2): 1179–94.

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*, Suntec City, Singapore, Association for Computational Linguistics, pp. 88–95.

Kim, H. D., and Zhai, C. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, New York, NY, USA, ACM, pp. 385–394.

Klosik, D. F., and Bornholdt, S. 2014. The citation wake of publications detects Nobel laureates' papers. *PLoS ONE* **9**(12): e113184. doi: 10.1371/journal.pone.0113184.

Kostoff, R. N., del Rio, J. A., Humenik, J. A., Garcia, E. O., and Ramirez, A. M. 2001. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology* **52**(13): 1148–56.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, pp. 282–89.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33**(1): 159–174.

Li, R., Chambers, T., Ding, Y., Zhang, G., and Meng, L. 2014. Patent citation analysis: calculating science linkage based on citing motivation. *Journal of the Association for Information Science and Technology* **65**(5): 1007–17.

Liu, J. S., Chen, H.-H., Ho, M. H.-C., and Li, Y.-C. 2014a. Citations with different levels of relevancy: tracing the main paths of legal opinions. *Journal of the Association for Information Science and Technology* **65**(12): 2479–88.

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., and Lin, Y. 2014b. Literature retrieval based on citation context. *Scientometrics* **101**(2): 1293–307.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Liu, Y., and Rousseau, R. 2014. Citation analysis and the development of science: a case study using articles by some Nobel prize winners. *Journal of the Association for Information Science and Technology* **65**(2): 281–9.

MacRoberts, M. H., and MacRoberts, B. R. 1984. The negational reference: Or the art of dissembling. *Social Studies of Science* **14**(1): 91–4.

Magerman, D. M. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, Cambridge, Massachusetts, Association for Computational Linguistics, pp. 276–83.

Milard, B. 2014. The social circles behind scientific references: relationships between citing and cited authors in chemistry publications. *Journal of the Association for Information Science and Technology* **65**(12): 2459–68.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., and Zajic, D. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, Boulder, Colorado, Association for Computational Linguistics, pp. 584–92.

Morante, R., and Blanco, E. 2012. *sem 2012 shared task: resolving the scope and focus of negation. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, Montréal, Canada, Association for Computational Linguistics, pp. 265–74.

Moravcsik, M. J., and Murugesan, P. 1975. Some results on the function and quality of citations. *Social Studies of Science* **5**: 86–92.

Nakov, P. I., Schwartz, A. S., and Hearst, M. A. 2004. Citances: citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, Sheffield, UK.

Nanba, H., Kando, N., and Okumura, M. 2004. Classification of research papers using citation links and citation types: towards automatic review article generation. In *Proceedings of the 11th SIG Classification Research Workshop*, Chicago, USA, pp. 117–34.

Nanba, H., and Okumura, M. 1999. Towards multi-paper summarization using reference information. In *IJCAI '99: Proceedings of the 16th International Joint Conference on Artificial Intelligence*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, pp. 926–31.

Nenkova, A., and Passonneau, R. 2004. Evaluating content selection in summarization: the pyramid method. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '04),* Boston, Massachusetts.

Och, F. J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* **29**(1): 19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pp. 311–18.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*, 2008, Marrakech, Morocco.

Prathap, G. 2014. A three-class, three-dimensional bibliometric performance indicator. *Journal of the Association for Information Science and Technology* **65**(7): 1506–8.

Qazvinian, V., and Radev, D. R. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, Coling 2008 Organizing Committee, 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 of the Association for Computational Linguistics*, Uppsala, Sweden, Association for Computational Linguistics, pp. 555–64.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Qazvinian, V., Radev, D. R., Mohammad S. M., Dorr, B., Zajic, D., Whidby, M., and Moon, T. 2013. Generating extractive summaries of scientific paradigms. *J. Artif. Int. Res.* **46**(1): 165–201. El Segundo, CA, USA.

Qazvinian, V., Radev, D. R., and Özgür, A. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23nd International Conference on Computational Linguistics (COLING-10)*, Beijing, China, pp. 895–903.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Radev, D. R., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. 2013. The acl anthology network corpus. *Language Resources and Evaluation* **47**(4): 919–44.

Radicchi, F., and Castellano, C. 2013. Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics* **97**(3): 627–37.

Rafols, I., and Meyer, M. 2009. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics* **82**(2): 263–87.

Shen, J., Yao, L., Li, Y., Clarke, M., Wang, L., and Li, D. 2013. Visualizing the history of evidence-based medicine: a bibliometric analysis. *Journal of the American Society for Information Science and Technology* **64**(10): 2157–72.

Small, H. 1982. Citation context analysis. In *Progress in Communication Sciences* **3**: 287–310.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning C. D., Ng, A. Y., and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington. Association for Computational Linguistics, pp. 1631–1642.

Spiegel-Rösing, I. 1977. Science studies: bibliometric and content analysis. *Social Studies of Science* **7**(1): 97–113.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday.

Swales, J. M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge Applied Linguistics. Cambridge, Cambridge University Press.

Teufel, S. 2006. Argumentative zoning for improved citation indexing. In J. G. Shanahan, Y. Qu, and J. Wiebe (eds.), Computing attitude and affect in text: Theory and Applications, **20**:159–69. Springer Netherlands. http://dx.doi.org/10.1007/1-4020-4102-0_13

Teufel, S., Siddharthan, A., and Tidhar, D. 2006. Automatic classification of citation function. In *Proceedings of EMNLP-06*, Sydney, Australia.

Thelwall, M., Haustein, S., Larivière, V., and Sugimoto, C. R. 2013. Do altmetrics work? twitter and ten other social web services. *PLoS ONE* **8**(5): e64841.

Thompson, G., and Yiyun, Y. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics* **12**(4): 365–82.

Velden, T., and Lagoze, C. 2013. The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology* **64**(12): 2405–27.

Vinkler, P. 2010. *The Evaluation of Research by Scientometric Indicators*, pp. 1–3. Chandos Learning and Teaching Series. Oxfordshire, United Kingdom: Chandos Publishing.

Waltman, L., van Eck, N. J., and Wouters, P. 2013. Counting publications and citations: Is more always better? *Journal of Informetrics* **7**(3): 635–41, ISSN 1751-1577, http://dx.doi.org/10.1016/j.joi.2013.04.001.

Wan, X., and Liu, F. 2014a. WL-index: leveraging citation mention number to quantify an individual's scientific impact. *Journal of the Association for Information Science and Technology* **65**(12): 2509–17.

Wan, X., and Liu, F. 2014b. Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology* **65**(9): 1929–38.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press

Weinstock, M. 1971. *Citation Indexes*, Kent, A. (ed.), vol. 5, Encyclopedia of Library and Information Science. New York: Marcel Dekker.

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

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. 2005. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, Vancouver, B.C., Canada, Association for Computational Linguistics, pp. 34–35.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, Cambridge, Massachusetts, Association for Computational Linguistics, pp. 189–96.

Yin, X., Huang, J. X., and Li, Z. 2011. Mining and modeling linkage information from citation context for improving biomedical literature retrieval. *Information Processing & Management* **47**(1): 53–67.

Zhang, C.-T. 2009. The e-index, complementing the h-index for excess citations. *PLoS ONE* **4**(5): e5429+.

Zhao, D., and Strotmann, A. 2014. In-text author citation analysis: feasibility, benefits, and limitations. *Journal of the Association for Information Science and Technology* **65**(11): 2348–58.

Ziman, J. M. 1968. *Public Knowledge: An Essay Concerning the Social Dimension of Science*. Cambridge, England, UK: Cambridge University Press.

Zitt, M., and Cointet, J.-P. 2013. Citation impacts revisited: how novel impact measures reflect interdisciplinarity and structural change at the local and global level. *ArXiv e-prints*.

https://doi.org/10.1017/S1351324915000443 Published online by Cambridge University Press