Measuring the Evolution of a Scientific Field through Citation Frames

David Jurgens
University of Michigan
jurgens@umich.edu

Srijan Kumar
Stanford University
srijan@stanford.edu

Raine Hoover
Stanford University
raine@stanford.edu

Dan McFarland
Stanford University
dmcfarla@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

Citations have long been used to characterize the state of a scientific field and to identify influential works. However, writers use citations for different purposes, and this varied purpose influences uptake by future scholars. Unfortunately, our understanding of how scholars use and frame citations has been limited to small-scale manual citation analysis of individual papers. We perform the largest behavioral study of citations to date, analyzing how scientific works frame their contributions through different types of citations and how this framing affects the field as a whole. We introduce a new dataset of nearly 2,000 citations annotated for their function, and use it to develop a state-of-the-art classifier and label the papers of an entire field: Natural Language Processing. We then show how differences in framing affect scientific uptake and reveal the evolution of the publication venues and the field as a whole. We demonstrate that authors are sensitive to discourse structure and publication venue when citing, and that how a paper frames its work through citations is predictive of the citation count it will receive. Finally, we use changes in citation framing to show that the field of NLP is undergoing a significant increase in consensus.

1 Introduction

Authors use citations to frame their contributions and connect to an intellectual lineage (Latour, 1987). An author's scientific frame employs citations in multiple ways (Figure 1) so as to build a strong

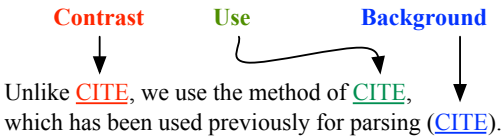

Figure 1: Examples of citation functionality.

and multifaceted argument. These differences in citations have been examined extensively within the context of a single paper (Swales, 1986; White, 2004; Ding et al., 2014). However, we know relatively little about how these citation frames develop over time within a field and what impact they have on scientific uptake.

Answering these questions has been largely hindered by the lack of a dataset showing how citations function at the field scale. Here, we perform the first field-scale study of citation framing by first developing a state-of-the-art method for automatically classifying citation function and then applying this method to an entire field's literature to quantify the effects and evolution of framing.

Analyzing large-scale changes in citation framing requires an accurate method for classifying the function a citation plays towards furthering an argument. Due to the difficulty of interpreting citation intent, many prior works performed manual analysis (Moravcsik and Murugesan, 1975; Swales, 1990; Harwood, 2009) and only recently have automated approaches been developed (Teufel et al., 2006b; Valenzuela et al., 2015). Here, we unify core aspects of several prior citation annotation schemes (White, 2004; Ding et al., 2014; Hernández-Alvarez and Gomez, 2016). Using this scheme, we create

Transactions of the Association for Computational Linguistics, vol. 6, pp. 391–406, 2018. Action Editor: Katrin Erk.
Submission batch: 8/2017; Revision batch: 12/2017; Published 6/2018.

one of the largest annotated corpora of citations and use it to train a high-accuracy method for automatically labeling a corpus. We apply our method to label the field of NLP, with over 134,127 citations in over 20,000 papers from nearly forty years of work.

Our work provides four key contributions for understanding how authors frame their citations. We introduce a new large-scale representative corpus of citation function and state-of-the-art methodology for classifying citations by function. We demonstrate that citations reflect the discourse structure of a paper but that this structure is significantly influenced by publication venue. Third, we show that differences in a paper's citation framing have a significant and meaningful impact on future scientific uptake as measured through future citations. Finally, by examining changes in the usage of citation functions, we show that the scholarly NLP community has evolved in how its authors frame their work, reflecting the maturation and growth of the field as a rapid discovery science (Collins, 1994). We publicly release our dataset and code to enable future research.

2 A Corpus for Citation Function

Citations play a key role in supporting authors' contributions throughout a scientific paper.[1] Multiple schemes have been proposed on how to classify these different roles, ranging from a handful of classes (Nanba and Okumura, 1999; Pham and Hoffmann, 2003) to twenty or more (Garfield, 1979; Garzone and Mercer, 2000). While suitable for expert manual analysis, many schemes include either fine-grained distinctions that are too rare to reliably identify or subjective classifications that require detailed knowledge of the field or author (Ziman, 1968; Swales, 1990; Harwood, 2009). Motivated by the desire to automatically examine large-scale trends in scholarly behavior, we address these issues by unifying the common aspects of multiple approaches in a single classification.

2.1 Classification Scheme

Our classification captures the broad thematic functions a citation can serve in the discourse, e.g., providing background or serving as contrast (Oppenheim and Renn, 1978; Spiegel-Rüsing, 1977; Teufel et al., 2006a; Garfield, 1979; Garzone and Mercer, 2000; Abu-Jbara et al., 2013).[2] Citation function reflects the specific purpose a citation plays with respect to the current paper's contributions. We unify the functional roles common in several classifications, e.g., (Spiegel-Rüsing, 1977; Garfield, 1979; Peritz, 1983; Teufel et al., 2006a; Harwood, 2009; Dong and Schäfer, 2011), into the six classes shown in Table 1, along with their description and example.

Our annotation scheme is similar to the six classes of Abu-Jbara et al. (2013) and the twelve-class scheme of Teufel (2000). The former has separate classes for comparison and for contrast, whereas the latter has multiple finer-grained distinctions for different kinds of comparison and contrasts. Here, we collapse these distinctions into a single class, COMPARISON AND CONTRAST, that signals the author is making some form of alignment between their work and another. In practice, we found that many citation contexts with alignments—such as this one—contain signals of both comparison and contrast; for our intended analyses, we considered this alignment signalling more important than whether the author was comparing or contrasting. Additionally, we introduce the FUTURE class to indicate that authors have forward-looking references for how their work might be applied later; these references are important for establishing a temporal lineage between works, and as we show later in §4, are the most frequent citation type in papers' Conclusion sections. Our adapted scheme enables us to conduct detailed analyses of the narrative structure of papers, venue citation pattern and evolution, and modeling the evolution of the whole field.

2.2 Annotation Process and Dataset

Annotation guidelines were created using a pilot study of 10 papers sampled from the ACL Anthology Reference Corpus (ARC) (Bird et al., 2008).

[1] For notational clarity, we use the term *reference* for the work that is cited and *citation* for the mention of it in the text.

[2] Another potential theme is citation sentiment (Athar, 2014; Kumar, 2016), but we omit this theme from our field-scale analysis because researchers have shown that negative sentiment is rare in practice (Chubin and Moitra, 1975; Vinkler, 1998; Case and Higgins, 2000) and can be quite subjective to classify due to textual mixtures of praise and criticism (Peritz, 1983; Swales, 1986; Brooks, 1986; Teufel, 2000).

Class	Description	Example
BACKGROUND	P provides relevant information for this domain.	This is often referred to as incorporating deterministic closure (Dörre, 1993).
MOTIVATION	P illustrates need for data, goals, methods, etc.	As shown in Meurers (1994), this is a well-motivated convention [...]
USES	Uses data, methods, etc., from P .	The head words can be automatically extracted [...] in the manner described by Magerman (1994).
EXTENSION	Extends P's data, methods, etc.	[...] we improve a two-dimensional multimodal version of LDA (Andrews et al, 2009) [...]
COMPARISON OR CONTRAST	Expresses similarity/differences to P.	Other approaches use less deep linguistic resources (e.g., POS-tags Stymne (2008)) [...]
FUTURE	P is a potential avenue for future work.	[...] but we plan to do so in the near future using the algorithm of Littlestone and Warmuth (1992).

Table 1: Our set of six functions a citation may serve with respect to a cited paper P.

Annotators completed two rounds of pre-annotation to discuss their process and design guidelines. All citations were then doubly-annotated by two trained annotators with expertise in NLP using the Brat tool (Stenetorp et al., 2012) and were then fully adjudicated to ensure quality. Following best practices for annotating citations (Athar, 2014), annotators saw an extended context before and after the citing sentence, provided from the output of ParsCit. Annotators were instructed to skip any instances whose context was corrupted or whose citance text did not match the regular citation style for ACL venues.[3]

The citation scheme was applied to a random sample of 52 papers drawn from the ARC. Each paper was processed using ParsCit (Councill et al., 2008) to extract citations and their references. As expected from prior studies (Teufel et al., 2006a; Dong and Schäfer, 2011), some citation functions were infrequent. We therefore attempted to oversample the infrequent classes FUTURE, EXTENSION, and MOTIVATION, by using keywords biased toward extracting citing sentences of a particular class (such as the word "future" for the FUTURE class). The resulting citing sentences were then annotated and could potentially be assigned to any class. In total, 1436 citations in context were annotated for the fully-labeled 52 papers (mean 27.6 citations/paper) and 533 supplemental contexts from 133 papers were added by targeted sampling, bringing the total number of instances to 1969. Table 2 shows the class distribution in the final dataset. Consistent with prior work, the majority of citations are BACKGROUND (Moravcsik and Murugesan, 1975; Spiegel-Rüsing, 1977; Teufel et al., 2006b).

Citation Function	Count
BACKGROUND	1021
USES	365
COMPARES OR CONTRASTS	344
MOTIVATION	98
CONTINUATION	73
FUTURE	68

Table 2: Citation class distribution in our dataset

3 Automatically Classifying Citations

The structure of a scientific article provides multiple cues for a citation's purpose. Our work draws on multiple approaches (Hernández-Alvarez and Gomez, 2016) to develop a classifier based on (1) *structural* features describing where the citation is located, (2) *lexical* and *grammatical* features for

[3] A small number of citation instances in our sample occurred in contexts where the surrounding text was malformed, which we attribute to being OCR errors, the citation being in the middle of a math-related context whose symbols were not converted, or where the citation occurred within a table or figure whose structure was treated as the surrounding text. In all cases, we viewed in the instance as unsuitable for use as a training example since it contained little meaningful context. These cases accounted for less than 10 instances in our data. A second set of instances were excluded when ParsCit mislabeled the span of a citation, either shortening it or increasing it to multiple citations' text. These wrong-spans occurred in less than 10 instances in our sample. A third set of citation instances were excluded due to citation style difference, where a paper in an earlier iteration of a conference used numeric citations, e.g., "[12]." These were excluded to ensure uniformity in the data and occurred in two papers that were excluded from in our initial sample. As these errors are sufficiently rare in our sample (<4%), we do not perform any further correction for these errors in the larger, un-annotated data.

Structural
section # and remaining # of sections
relative positions in paper, section, subsection sentence, & clause
of other citations in subsection, sentence, & clause
canonicalized section title
Lexical, Morphological, and Grammatical
function patterns of Teufel (2000)
topical similarity with cited paper
the presence of each of 23 connective phrases
verb tense
lengths of the containing sentence and clause
whether used inside of a parenthetical statement
† bootstrapped function patterns
† custom function patterns
† citation prototypicality
† citation context topics
† paper topics
† whether used in nominative or parenthetical form,
† whether preceded by a Pascal-cased word
† whether preceded by an all-capital case word
Field
of years difference in publication dates
whether the cited paper is a self-citation
† citing paper's venue: journal/conference/workshop
† reference's venue: journal/conference/workshop
reference's citation count, and PageRank (at time of the citation)
† reference's Hub and Authority scores and Network Centrality (at time of the citation)
† # of citations in common
Usage
of indirect citations
of direct citations
of indirect citations per section type
of direct citations per section type
fraction of bibliography used by this reference

Table 3: Features for classifying citations. Novel features are marked with a †.

how the citation is described, (3) *field* features that take into account venue or other external information, and (4) *usage* features on how the reference is cited throughout the paper. Table 3 shows our features, which includes ten novel feature types, in addition to several drawn from recent systems (Teufel, 2000; Teufel et al., 2006b; Dong and Schäfer, 2011; Wan and Liu, 2014; Valenzuela et al., 2015; Zhu et al., 2015).

Function	Pattern
COMP. OR CON.	@SIMILAR_ADJ to @REFERENTIAL @USE
COMP. OR CON.	the @RESEARCH_NOUN of #N
EXTENDS	@CHANGE_NOUN of #N 's
EXTENDS	@CHANGE_NOUN of *citation* 's
MOTIVATION	@INSPIRATION by #N
USES	@1ST_PERSON_PRONOUN_(NOM) @USE the #N
USES	the #N corpus
USES	#D #N #N *citation*

Table 4: Examples of bootstrapped patterns learned and their associated class where @ denotes a lexical class and # denotes a part of speech wild card.

3.1 Features

Following, we describe in detail the three main categories of novel features.

Pattern-based Features Patterns provide a powerful mechanism for capturing regularity in citation usage (Dong and Schäfer, 2011). Our patterns are a sequence of cue phrases, parts of speech, or lexical categories, like positive-sentiment words or specific categories that allow generalizations across phrases like "we extend" and "we build upon." We began with the largest publicly-available list of citation patterns (Teufel, 2000) and extended it with 132 new patterns and 13 new lexical categories based on a manual analysis of the corpus.

We then used bootstrapping to automatically identify new patterns as follows: Each annotated context was converted into fixed-length patterns using (a) our 42 lexical categories, (b) part of speech wild cards, or (c) the tokens directly. To avoid semantic drift (Riloff and Jones, 1999), a bootstrapped pattern was only included as a feature if the majority of its occurrences were with a single citation function.[4] Table 4 shows examples of these bootstrapped patterns.

Previous patterns primarily use cues from the same sentence as the citation (Teufel, 2000). However, authors often use multiple sentences to indicate a citation's purpose (Abu-Jbara and Radev, 2012; Ritchie et al., 2008; He et al., 2011; Kataria et al., 2011). For example, authors may first introduce a work positively, only to contrast with it in later sen-

[4]For computational efficiency, patterns were restricted to having between 3 and 8 tokens and at most two part of speech wild cards. Due to its high frequency, patterns for BACKGROUND were required to occur in at least 100 contexts.

1) algorithm parameter model training method clustering
2) measure score metric information similarity distance
3) % result accuracy report achieve performance system
4) training weight feature och model set algorithm error
5) work related previous paper problem approach present

Table 5: The most probable words from five example topics learned from citation contexts.

tences (Peritz, 1983; Brooks, 1986; Mercer et al., 2004). Indeed the average text pertaining to a citation spans 1.6 sentences in the ARC (Small, 2011).

We therefore induce bootstrapped patterns specific to the citation sentence as well as the preceding and following sentences. Ultimately, 805 new bootstrapped patterns were added for the citing sentence, 669 for the preceding context, and 1159 for the following context, a total of over four times the number of manually curated patterns.

Topic-based Features A context's thematic framing can point to the purpose of a citation even in the absence of explicit cues. For example, a citation in a context describing system performances and results is likely to be a COMPARE OR CONTRAST, whereas one describing methodology is more likely to be USES. We quantify this thematic framing by using features based on topic models, computed over the sentence containing the citation and also over the paragraph containing the citing sentence. For each type of context, a topic model is trained over 321,129 respective contexts from the ARC. Table 5 shows example topics.

Prototypical Argument Features We also explored richer grammatical features, drawing on selectional preferences reflecting expectations for predicate arguments (Erk, 2007). We construct a prototype for each citation function by identifying the frequent arguments seen in different syntactic positions. For example, EXTENDS citations occur frequently as objects of verbs such as "follow" and "use", whereas USES citations have techniques or artifact words as dependents; Table 6 shows more examples. Each class's selectional preferences are represented using a vector for the argument at each relation type, constructed by summing the vectors of all words appearing in it. Each function is represented as a separate feature whose value is the average similarity of an instance's arguments with the class's preferences for all observed syntactic relationships (i.e., how similar are the syntactically-related words to the function's preferences). Our work differs from dependency-based features from prior work that use separate features for each unique dependency path and argument (Athar and Teufel, 2012; Abu-Jbara et al., 2013); in contrast, we use a single feature for each path with distributed representation for its arguments, which allows our features to generalize to similar words that are unseen in the training data.

Function	Path	Arguments
MOTIVATION	nmod^{-1}	inspire, work, show
MOTIVATION	nmod^{-1}, nmod^{-1}	exemplify, direction, inspire
MOTIVATION	nsubj^{-1}	show, use, suggest
USES	nmod^{-1}	use, describe, propose
USES	dobj	use, follow, see
USES	dep^{-1}	system, algorithm, mechanism
COMP. OR CONT.	nmod^{-1}, nmod^{-1}	similar, related, use
COMP. OR CONT.	dep^{-1}	system, method, approach
COMP. OR CONT.	nsubj^{-1}, dobj	approach, technique, rule
EXTENDS	amod	previous, prior, unsupervised
EXTENDS	nmod^{-1}, nmod^{-1}	base, version, extension
EXTENDS	dobj^{-1}	follow, extend, unfold

Table 6: Examples of citation function selectional preferences with the most-frequent arguments seen for each paths. Each dependency path feature value reflects the similarity of (i) the average word vector for that path's arguments with (ii) the vector of the path's argument in a given context, if the path is present.

3.2 Experimental Setup

Models All models were trained using a Random Forest classifier, which is robust to overfitting even with large numbers of features (Fernández-Delgado et al., 2014). After limited grid search over possible configurations,[5] we set parameter values as follows. The number of random trees is 2500 and we required each leaf to match at least 5 instances. To overcome the class imbalance, we use SMOTE (Chawla et al., 2002) to generate synthetic examples in the training fold using the 5 nearest neighbors. The

[5]The grid search was performed using the following parameter ranges: number of trees $[100, 500, 1000, 2500]$; maximum number of depth of the decision tree as $\frac{n}{10}$ or $\sqrt{n}$, where n is the number of features; minimum leaf size in decision tree $[2, \ldots, 10]$; number of topics $[50, 100, 250, 500]$; and whether to use Smote (Chawla et al., 2002).

classifier is implemented using SciKit (Pedregosa et al., 2011) and syntactic processing was done using CoreNLP (Manning et al., 2014). Selectional preferences used pretrained 300-dimensional GloVe vectors from the 840B token Common Crawl (Pennington et al., 2014). The topic model features used an LDA with 100 topics.

Data Annotated data is crucial for developing high accuracy for rare citation classes. Therefore, we integrate portions of the dataset of Teufel (2010),[6] which has fine-grained citation function labeled for ACL-related documents using the annotation scheme of Teufel et al. (2006b). We map their 12 function classes into our six classes (see Appendix A). When combining the two datasets, we omit the data labeled with their BACKGROUND-equivalent class to reduce the effects of a large majority class and because instances of the FUTURE class are merged into BACKGROUND according to their scheme. The resulting citation function dataset contains 3,083 instances.

Evaluation Evaluation is performed using cross-validation where each fold leaves out all citations of a single paper. Stratifying by paper instead of instance is critical: since multiple citations may appear in the same sentence, instance-based stratification would leak information between training and test. We also note that when performing cross-validation, we compute the bootstrapped patterns and prototypical argument features using only contexts from the training data. We report macro-averaged F1 scores across the six function classes.

Comparison Systems We compare against three state-of-the-art systems which all use similar citation function classifications. Abu-Jbara et al. (2013) use a combination of lexicons, structural, and syntactic features for classification. Instances are classified using a linear kernel SVM. Their described method also uses a second CRF-based classifier to include neighboring sentences in the citation context. As the dataset for this citation-span classifier is not public, we are unable to reproduce this part of their system. However, the authors note that using the citing sentence alone is the correct context in 80% of the instances, so we view our implementation as a close approximation. Dong and Schäfer (2011) classify citations using a small set of lexicons and discourse features, which includes regular expressions on sentence parts of speech for capturing syntactic cues. Their model uses a naive bayes classifier, which was shown to work well for their data. Teufel (2000) is the most similar model to ours as it uses a subset of our lexical features and lexicons; the model uses a k-nearest neighbor classifier. We note that the original implementation used a custom syntactic tool for identifying aspects like verb tense, which we replaced with CoreNLP. We compare against the system Teufel (2000) instead of the system Teufel et al. (2006b) because the latter includes pattern-based features that are not fully specified or publicly available; however, the two systems are similar in their description. For all three compared systems, we use identical parameter values as reported in the papers.

[6]Their original data may be obtained at `http://www.cl.cam.ac.uk/~sht25/CFC.html` and we distribute a re-annotated version of this with our data.

System	Macro F1
This work	**0.530**
without topic features	0.502
without selectional prefs.	0.464
without bootstrapped pats.	0.457
without any novel features	0.474
Abu-Jbara et al. (2013)	0.410
Teufel (2000)	0.273
Dong and Schäfer (2011)	0.233
Majority-Class	0.092
Random	0.138

Table 7: Classifier performances.

Baselines Two baselines are used for comparison: a Random baseline that selects a function at chance and a Majority-class baseline that labels all instances with the most frequent citation function BACKGROUND.

3.3 Results and Discussion

Our methods substantially outperformed the closest state of the art and both baselines for both classification tasks, as shown in Table 7. All improvements over comparison systems are statistically significant (McNemar's, $p \leq 0.01$). The closest-performing system was that of Abu-Jbara et al. (2013), which also had a heavily-lexicon based approach.

An ablation test suggests that each of our novel

features contributed to the final performance. Notably, we observe that selectional preference and bootstrapped lexicon features had the largest impact on performance; both features capture local information indicating this type of information is important for recognizing function. While multiple prior works have focused on patterns to recognize function, our results suggest that machine learned patterns and contextual regularities (topics or word vectors) provide highly-accurate information. Indeed, examining the feature weighting in the random forest shows that features for structure (e.g., section number), topic, and selectional preference comprised most of the 100 highest-weighted features (76%).

The use of conjunctive features by the Random Forest was critical for high performance. All other non-conjunctive classifiers we tried resulted in substantially lower Macro F1: Naive Bayes, 0.286; k-nearest neighbor, 0.255 (k=3); and Linear-kernel SVM, 0.393 (C=1).[7]

The resulting classifier performance is sufficient to apply it to the entire ARC dataset for the analyses in the next four sections. Nonetheless, errors remain. Our error analysis revealed that a main challenge is incorporating information external to the citing sentence. Consider the following example:

> BilderNetle is our new data set of German noun-to-ImageNet synset mappings. ImageNet is a large-scale and widely used image database, built on top of WordNet, which maps words into groups of images, called synsets (Deng et al., 2009).

Here the citing sentence appears much like a BACKGROUND citation when read in isolation; however, the preceding sentence reveals that the citing work's data is based on the citation, making its function USES though no explicit cues suggest this in the citing sentence. Thus, our error analysis supports the observation of Abu-Jbara et al. (2013) that citation context identification is an important step towards improving performance and models with richer textual understanding are needed to understand how the citation relates to the broader context and narrative outside of the sentence.

[7]We observed mixed results when using a random forest with other approaches. Replacing the k-nearest neighbors classifier used in Teufel (2000) with a random forest improves citation function classification by 0.119 Macro F1. In contrast, replacing the SVM model used by Abu-Jbara et al. (2013) decreased performance by 0.072 Macro F1. We speculate that the larger feature space of Teufel (2000), which is more similar to our features space, is more conducive to conjunctive features.

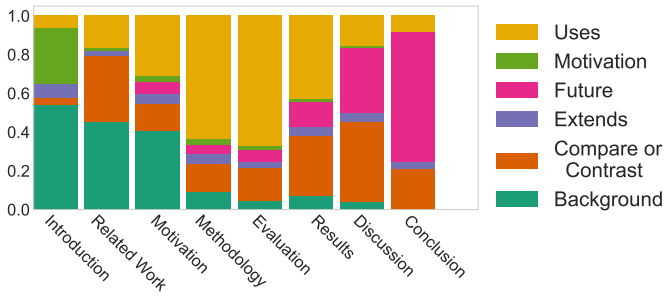

Figure 2: Expected percentage of citation functions per section shows a clear narrative trajectory across sections.

In the next four sections, we apply our classifier trained on our combined dataset (2600 citation instances) to the ACL Anthology to study what citation functions can tell us about scientific uptake and author behavior.

4 Narrative Structure of Citation Function

Scientific papers commonly follow a structured section narrative to frame their contributions: Introduction, Methodology, Results, and Discussion (Skelton, 1994; Nwogu, 1997). Each section in the narrative adopts argumentative moves designed to convince the reader of the work's claims (Swales, 1986; Swales, 1990). We hypothesize that this narrative is mirrored in how authors use their citations in sections, with the citation's function serving to further evoke section's intended rhetorical frame (Goffman, 1974; Gumperz, 1982).

To test this hypothesis, the function classifier was applied to all 21,474 papers of the latest 2016 release of the ACL Anthology. This yielded a dataset of 134,127 citations between papers in the ARC. The resulting distributions of citation function (Figure 2), show that authors' citation framing indeed parallels the expected rhetorical framing seen in the writing: (1) establishing an intellectual lineage via BACKGROUND citations in the Introduction, Motivation, and Related Work sections to (2) introducing methodology with USES citations in the Methodology and Evaluation sections, (3) a large increase in COMPARISON OR CONTRAST for related literature in the Results and Discussions, and finally (4) closing comparisons and pointers to future directions.

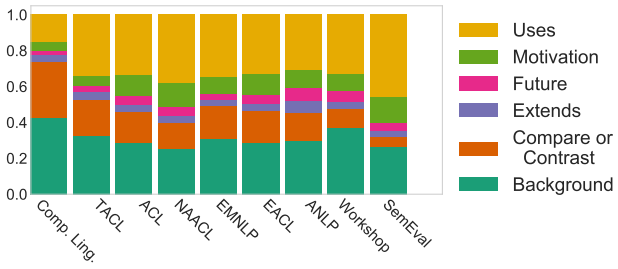

Figure 3: Venues attract different citation framing, as seen in the differences in the distribution of citation functions per venue in the journals (left two), workshops (right two) and conferences (middle).

These trends also mirror the thematic structure identified in full-paper textual analyses (Skelton, 1994; Nwogu, 1997). By showing that a section contains citations serving a variety of functions, our findings further point to a new direction for citation placement studies (Hu et al., 2013; Ding et al., 2013; Bertin et al., 2016), which have largely treated all citations within a section as equivalent.

5 Venues and Citation Patterns

Each publication venue has its own expectation for the types of work it accepts, e.g., the degree of polish or depth of experiments. As such, each venue has a distinct genre of writing, from the tentative results of workshop papers to journal papers with substantial synthesis. To what degree do venue genres affect the way authors cite? To answer this, we used the same experimental setup as the previous section. Figure 3 shows citation function by venue for the 134,127 citations.

We find that similar venue types have similar distributions of citation framing. Journals have the highest percentage of BACKGROUND citations, suggesting that their extra space and wider temporal scope lends itself more to positioning. Conference venues devote proportionally more space to contrast and comparison with other work, presumably because new NLP work is first presented at conferences and hence acceptance requires demonstrating the proposed technique is better than existing ones. Workshops, by contrast, have relatively little comparison and instead use more BACKGROUND; the experimental nature of workshop papers presumably results in fewer potential prospects for comparison. Similarly, the SemEval workshops focus on rapidly developing new systems for a shared task, which is reflected in the papers framing as primarily USES citations and relatively little COMPARE OR CONTRAST, as the venue's shared task provides the broader framing connecting papers to related work.

6 Venue Evolution

The growth of the ACL community has been accompanied by the creation of new publication venues. How have these new venues evolved by possibly becoming institutionalized and resembling established conferences or becoming stylistically distinct and capturing different representations of knowledge? Citation framing provides an ideal lens for observing this evolution by measuring the degree to which a newer venue's papers' framing mirrors that used by papers in established venues. Here, we examine venue evolution in the ACL through its workshops.

Conferences within the ACL community frequently have collocated workshops that focus on a particular theme and have their own proceedings. The number of workshops has increased substantially with the growth of the field, from around ten workshops in the 1990s to over 100 workshops by 2010, with many workshops having multiple iterations across the years. This growth has led to the observation that ACL workshops have become like mini-conferences rather than venues for early-stage research and discussion (Daumé III, 2016). Are workshops becoming more conference-like and, if so, is this a general trend or primarily seen in long-running workshops? We hypothesize that multiple iterations of the same workshop create institutional knowledge and community norms that leads to more conference-like papers over time. Here, we test this hypothesis by measuring whether workshop papers have become more similar in their citation framing to papers from the main conferences.

Experimental Setup We repeat the classification setup from the previous experiment. We compare the average framing of a paper within a venue in a given year with the distribution for the two main conferences (ACL and NAACL) within that year. Distributions are compared using the Jensen-Shannon Divergence, where 1 indicates that the venues are citing identically.

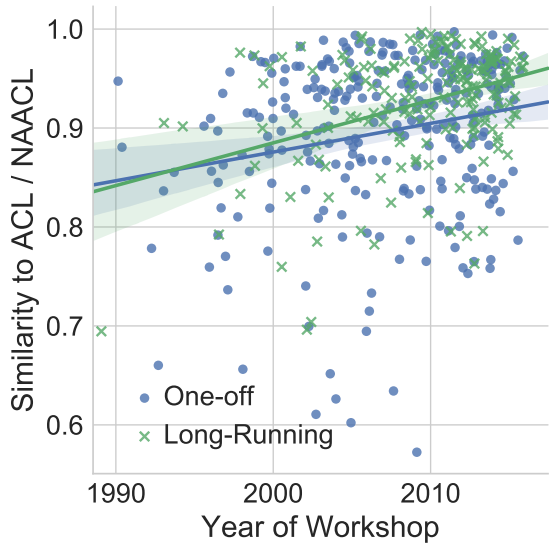

Figure 4: Both one-off and long-running workshops have proceedings that increasingly appear more conference-like in how their papers frame their citations, suggesting that workshops are in fact becoming more conference-like. Lines show a fitted linear regression with bootstrapped 95% confidence intervals, which stop overlapping beginning in 2007.

Results Workshops consistently became more conference-like in their papers' citation framing (Figure 4). Further, this trend in increasing similarity was seen much more for both long-running workshops, suggesting that multiple-iteration workshops create their own conference-like norms that attract more conference-like papers each year. We speculate that the increasing similarity of workshops to conferences indicates the field has begun to congeal. Early workshops were like satellite conferences on peripheral topics, but as the field grows and the methods standardize, a norm of publication emerges such that conferences and workshops all resemble an institutionalized standard. Multiple iterations of a work accelerate this process by further establishing publication norms within a sub-community.

7 Predicting Future Impact

The scholarly narrative told through citations provides the reader with support for its claims and technical competence (Latour, 1987, p. 34). This framing could affect how the work is perceived and, ultimately, how it is received and cited within the community (Shi et al., 2010). Does the frame evoked by a paper through its citation functions (the way it compares to related work, or motivates, or points to the future) affect its reception?

Experimental Setup To quantify how a paper's citation framing affects its future uptake, we constructed a negative binomial regression to predict the cumulative number of citations a paper received within the first five years after publication, which is known to be highly representative of the eventual citation count (Wang et al., 2013; Stern, 2014). In addition to variables for how the paper cites, we include variables from state-of-the-art features for predicting the citation count (Yan et al., 2011; Yogatama et al., 2011; Yan et al., 2012; Chakraborty et al., 2014; Dong et al., 2016), described below. We compare against a baseline regression model without citation framing and test whether the model's fit is improved when the framing is included as features.

All papers with at least five years of publication history in the anthology were considered, yielding a set of 10,434 papers. We used negative binomial models, which are more appropriate than linear regression as citation counts are non-negative discrete counts, and compared them by using Akaike Information Criterion (AIC). AIC measures each model's goodness of fit in proportion to the number of independent variables; when comparing models, the model with the minimal AIC is preferred (Akaike, 1974). If citation framing helps to explain future impact, we should see a lower AIC despite the penalty for including more variables to the model.

Five types of non-citation features were included. To model the amount of attention received by different research areas, each paper is associated with its distribution over 100 topics, built using LDA over the ARC. To capture diversity, we include the entropy of the topic distribution. We include the publication year since the size of the field changes over time. Multi-author papers are known to receive higher citation counts (Gazni and Didegah, 2011), partially due to the effects of self-citation (Fowler and Aksnes, 2007), and therefore we include the number of authors on the paper. To reflect how integrated the paper is, we include the number of references.[8]

[8]To control for collinearity between citation-related predic-

	Baseline	with Framing
Intercept	-173.334^{***}	-161.375^{***}
# of authors	0.101^{***}	0.101^{***}
# of citations	0.037^{***}	0.036^{***}
year	0.088^{***}	0.082^{***}
topic diversity	-0.741^{***}	0.685^{***}
BACKGROUND		0.013^{**}
COMP. OR CON.		0.025^{***}
EXTENDS		0.021^{**}
FUTURE		0.016
MOTIVATION		0.014^{*}
USES		0.055^{***}
Log Likelihood	−17,485.700	−17,416.820
Akaike Inf. Crit.	35,693.400	35,567.630

$^{*}\ p < 0.1, ^{**}\ p < 0.05, ^{***}\ (p < 0.01$

Table 8: Regression models for predicting the total number of citations five years after publication show that a paper's citation framing provides a statistically significant improvement in model fit and reveals which type of framing yields more cited papers. Regression coefficients for venue and topics are omitted for space.

Results Knowledge of how a paper frames its contributions helps improve predicting its future impact, with a statistically significant improvement in AIC when the distribution of citation functions is added (likelihood ratio test, $p \leq 0.01$).

Table 8 shows which types of citations are significantly predictive of higher impact ($p \leq 0.01$). Two main insights can be made from these results. First, papers maximize their future impact when framed as integrating many other technologies via USES citations. Second, works that frame their contributions through COMPARISON OR CONTRAST rather than BACKGROUND are more likely to have a higher impact. Latour (1987, p. 54) has suggested that authors may deflect criticism of their work (improving its perception) by claiming it as an extension, rather than comparing it with prior work. However, we did not observe this effect in how authors frame their work as COMPARISON OR CONTRAST or EXTENDS, with both having significant positive effects.

tors, we regress out the number of citations from the citation function counts (Kutner et al., 2004; O'brien, 2007). Finally, we include the publication venue, using the individual conference or workshop in which the paper was published to control for variations in prestige between venues. The resulting model has a variance inflation factor of < 10 for all variables.

8 The Growth of Rapid Discovery Science

As scientific fields evolve, new subfields initially emerge around methods or technologies which become a focus of collective puzzle-solving and continual improvement (Moody, 2004). NLP has witnessed the emergence of several such subfields from the early grammar based approaches in the 1950s-1970s, to the statistical revolution in the 1990s, to the recent deep learning models (Spärck Jones, 2001; Anderson et al., 2012). Collins (1994) proposed that a field can undergo a particular shift, referring to it as *rapid discovery science*, when the field (a) reaches high consensus on research topics as well as methods and technologies, and (b) then develops genealogies of methods and technologies that continually improve on one another. Over time, there is increased consensus on core approaches, and the field's periphery is extended to new research puzzles rather than contesting prior efforts. Collins claims this shift characterizes natural sciences, but not many social sciences, which are instead more likely to engage in continual contesting and turnover of core methods and assumptions (Evans et al., 2016).

We argue that a shift to rapid discovery science should be visible in the way citations are used to frame works in the field as a whole. Specifically, we expect that as consensus is reached (1) authors are expected to have fewer comparisons to other works and instead can simply acknowledge past work as background and (2) the remaining comparisons concentrate on fewer works, reflecting those works' status as accepted benchmarks of performance. Further, we expect that as a methodological lineage develops we should also observe an increased concentration of USES citations on papers describing methods and data.

We propose that the increased use of shared evaluations, and the statistical methodology borrowed originally from electrical engineering (Hall et al., 2008; Anderson et al., 2012) has led NLP to undergo a shift towards rapid discovery science.

Experimental Setup We repeat the setup of previous experiments and measure the expected citation

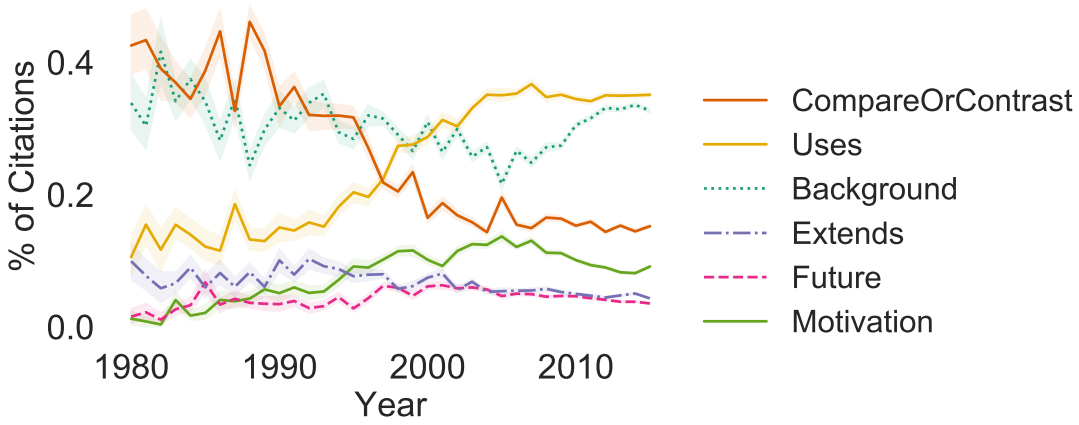

Figure 5: Changes in the average citation frame in ACL papers reveals a continued decline in the percentage of COMPARISON OR CONTRAST and increase in USES citations. The increase in BACKGROUND citations circa 2010 marks the start of the era of unlimited references in ACL conferences. Shaded regions show bootstrapped 95% confidence intervals.

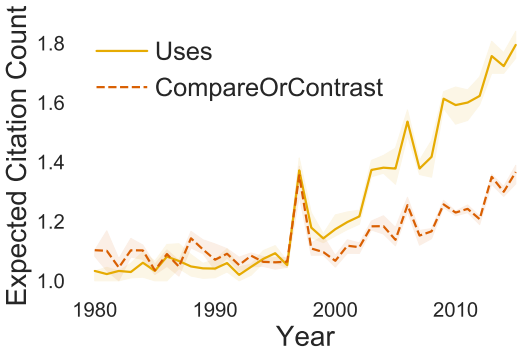

Figure 6: The average cited paper receives an increasing number of USES and COMPARISON OR CONTRAST citations per year showing that field increasingly builds upon the same set of papers, providing a methodological lineage. Shaded regions show bootstrapped 95% confidence intervals.

frame of a paper per year using all papers published in that year.

Results The NLP field shows a significant increase in consensus consistent with the rise in rapid discovery science, evidenced through two main trends.

First, NLP authors use a decreasing number of comparison and contrast citations (r= -0.899, p $\leq$ 0.01) as seen in Figure 5. Instead of comparing to others, it seems that authors simply acknowledge prior work as BACKGROUND, which had a corresponding increase in relative frequency. Despite an increase in BACKGROUND citations, the total percentage of non-methodological citations still declines (r= -0.663, p $\leq$ 0.01), with authors instead increasingly including more USES citations. Latour (1987, p. 50) argues that such non-methodological references are critical to an author's defense of an idea. We therefore interpret the observed decrease in non-methodological references as signaling a reduced need for authors to defend aspects of their work. Authors are able to compare against fewer papers due to the field's growing consensus on the validity of the problem and methodological contribution.

Note that there is a small but significant increase in the number of non-methodological references between 2009 and 2011. This transition corresponds to the date at which ACL venues began allowing unlimited references (2010 for ACL, 2011 for NAACL, etc.). Unlimited extra space for citations acted to modify authors' citation framing behavior; given unlimited space, authors chose to include proportionally more non-methodological citations.[9]

In the second trend, authors are more likely to use and compare against the same set of papers, as shown in Figure 6 by the rise in expected incom-

[9]Note that this change acts against the general decrease in non-methodological; considering only 1980-2009, the decrease in non-methodological is even larger (r= -0.568, p $\leq$ 0.01).

ing citations to those works compared against (r= 0.734, p $\leq$ 0.01) and used (r=0.889, p $\leq$ 0.01). For example, in 1991, authors compared with a diffuse group of parsing papers, e.g., (Shieber, 1988; Pereira and Warren, 1983; Haas, 1989), with such papers receiving at most three citations that year; whereas in 2000, most comparisons were to a core set of parsing papers, e.g., (Collins, 1999; Buchholz et al., 1999; Collins, 1997), with a much sharper (lower entropy) distribution of citations. These trends show the increased incorporation of prior work to form a lineage of method technologies as well as show increased consensus on which works are sufficient for comparing against in order to establish a claim. These results also empirically confirm the observation of Spärck Jones (2001) that a major trend in NLP in the 1990s was an increase in reusable technologies and evaluations, like the BNC (Leech, 1992) and the Penn Treebank (Marcus et al., 1993).

More broadly, our work points to the future of NLP as a quickly moving field of high consensus and suggests that artifacts that facilitate consensus such as shared tasks and open source research software will be necessary to continue this trend.

9 Conclusion

Authors cite works for different reasons (or function), so regarding them as equivalent signals is potentially problematic. Many fluff citations exist, while some less common ones are substantively relevant to the paper's argument. A careful analysis of citation reveals that authors cite works for multiple reasons—as background, motivation, extension, use, contrast, or future. When authors utilize some forms of citation over others they can significantly influence how their own work gets perceived and taken up by others (Latour, 1987). Simply put, citation functions help frame an article's reception. Moreover, a differentiation of citation functions affords a deeper understanding of how scholars develop arguments for different publication venues as well as how these venues may demand different forms of knowledge representation and arguments over time. In fact, these modes of citation help us understand the state of research efforts and their evolution more broadly for entire scientific fields like NLP. In this paper, we relate all this using a new corpus annotated with citation function and by developing a state-of-the-art classifier for revealing scientific framing. In doing so, we demonstrate the importance of novel unsupervised features related to topic models and argument structure, and label all the citations for an entire field.

We then show that citation framing reveals salient behaviors of writers, readers, and the field as a whole: (1) authors are sensitive to discourse structure and venue when citing, (2) ACL workshops have evolved to become more like the mainstream conferences, with multi-iteration workshops being quicker to establish conference-like norms, (3) the way in which an author frames their work aids in predicting its future impact as the number of citations its receives, with the community favoring works that integrate many new technologies and also relate to prior work through comparison and contrast, and (4) the NLP field as a whole has seen increased consensus in what constitutes valid work—with a reduced need for positioning and excessive comparison—demonstrating its shift towards rapid discovery science. All data, materials, and code for all systems are available at `https://github.com/davidjurgens/citation-function`.

Acknowledgements

The authors thank Jure Leskovec, Vinod Prabhakaran, Will Hamilton, and the other members of the Stanford NLP Group for helpful discussions and comments and thank Min-Yen Kan for hosting the ACL Anthology and help with data. We also thank the area chair, Katrin Erk, and reviewers for their helpful comments and suggestions. This work is also partially supported by the NSF under award IIS-1633036, the Stanford Data Science Initiative, the Brown Institute for Media Innovation, and the Science Surveyor project.

References

Amjad Abu-Jbara and Dragomir Radev. 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 (NAACL-HLT)*, pages 80–90. Association for Computational Linguistics.

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R. Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 596–606.

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a computational history of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601. Association for Computational Linguistics.

Awais Athar. 2014. Sentiment analysis of scientific citations. *Technical Report, University of Cambridge, Computer Laboratory*.

Marc Bertin, Iana Atanassova, Yves Gingras, and Vincent Larivière. 2016. The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology (JASIST)*, 67(1):164–177.

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan R. Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

Terrence A. Brooks. 1986. Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37(1):34–36.

Sabine Buchholz, Jorn Veenstra, and Walter Daelemans. 1999. Cascaded grammatical relation assignment. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 239–246.

Donald O. Case and Georgeann M. Higgins. 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7):635–645.

Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2014. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th Joint Conference on Digital Libraries (JCDL)*, pages 351–360.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Dale E. Chubin and Soumyo D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5:423–441.

Randall Collins. 1994. Why the social sciences won't become high-consensus, rapid-discovery science. *Sociological Forum*, 9(2):155–177.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 16–23.

Michael Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania.

Isaac G. Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

Hal Daumé III. 2016. Workshops and mini-conferences, November. `https://nlpers.blogspot.com/2016/11/workshops-and-mini-conferences.html`.

Ying Ding, Xiaozhong Liu, Chun Guo, and Blaise Cronin. 2013. The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, 7(3):583–592.

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology (JASIST)*, 65(9):1820–1833.

Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 623–631.

Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. 2016. Can scientific impact be predicted? *IEEE Transactions on Big Data (TBD)*, 2(1):18–30.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 216–223.

Eliza D. Evans, Charles J. Gomez, and Daniel A. McFarland. 2016. Measuring paradigmaticness of disciplines using text. *Sociological Science*, 3:757–778.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research (JMLR)*, 15(1):3133–3181.

James H. Fowler and Dag W. Aksnes. 2007. Does self-citation pay? *Scientometrics*, 72(3):427–437.

Eugene Garfield. 1979. *Citation Indexing Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons, New York.

Mark Garzone and Robert E. Mercer. 2000. Towards an automated citation classifier. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, pages 337–346.

Ali Gazni and Fereshteh Didegah. 2011. Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. *Scientometrics*, 87(2):251–265.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience.* Harvard University Press.

John J. Gumperz. 1982. *Discourse strategies*, volume 1. Cambridge University Press.

Andrew Haas. 1989. A parsing algorithm for unification grammar. *Computational Linguistics*, 15(4):219–232.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 363–371.

Nigel Harwood. 2009. An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, 41(3):497–518.

Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 755–764. ACM.

Myriam Hernández-Alvarez and Josém Gomez. 2016. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, 22(03):327–349.

Zhigang Hu, Chaomei Chen, and Zeyuan Liu. 2013. Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, 7(4):887–896.

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and C. Lee Giles. 2011. Context sensitive topic models for author influence in document networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 22, page 2274. Citeseer.

Srijan Kumar. 2016. Structure and dynamics of signed citation networks. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW)*.

Michael H. Kutner, Chris Nachtsheim, and John Neter. 2004. *Applied Linear Regression Models*. McGraw-Hill/Irwin.

Bruno Latour. 1987. *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Geoffrey Leech. 1992. 100 million words of English: The British National Corpus (BNC). *Language Research*, 28(1):1–13.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the System Demonstrations at 52th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 55–60.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Robert E. Mercer, Chrysanne Di Marco, and Frederick W. Kroon. 2004. The frequency of hedging cues in citation contexts in scientific writing. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 75–88. Springer.

James Moody. 2004. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2):213–238.

Michael J. Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5(1):86–92.

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 926–931.

Kevin Ngozi Nwogu. 1997. The medical research paper: Structure and functions. *English for Specific Purposes*, 16(2):119–138.

Robert M. O'brien. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

Charles Oppenheim and Susan P. Renn. 1978. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, 29(5):227–231.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research (JMLR)*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Fernando C.N. Pereira and David H.D. Warren. 1983. Parsing as deduction. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics (ACL)*, pages 137–144.

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

S. B. Pham and A. Hoffmann. 2003. A new approach for scientific citation classification using cue phrases. In *Proceedings of the Australian Joint Conference in Artificial Intelligence (AI)*, pages 759–771.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*, pages 474–479.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 213–222.

Xiaolin Shi, Jure Leskovec, and Daniel A. McFarland. 2010. Citing for high impact. In *Proceedings of the 10th annual Joint Conference on Digital libraries (JCDL)*, pages 49–58. ACM/IEEE-CS.

Stuart M Shieber. 1988. A uniform architecture for parsing and generation. In *Proceedings of the 12th Conference on Computational Linguistics*, pages 614–619. Association for Computational Linguistics.

John Skelton. 1994. Analysis of the structure of original research papers: An aid to writing original papers for publication. *British Journal of General Practice*, 44(387):455–459.

Henry Small. 2011. Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics*, 87(2):373–388.

Karen Spärck Jones. 2001. Natural language processing: a historical review. *University of Cambridge*, pages 2–10.

Ina Spiegel-Rüsing. 1977. Bibliometric and content analysis. *Social Studies of Science*, 7:97–113.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 102–107.

David I. Stern. 2014. High-ranked social science journal articles can be identified from early citation information. *PloS One*, 9(11):e112520.

John Swales. 1986. Citation analysis and discourse analysis. *Applied linguistics*, 7(1):39–56.

John Swales. 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*. Cambridge University Press, Cambridge, UK.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006a. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 80–87. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006b. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 103–110.

Simone Teufel. 2000. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Ediburgh.

Simone Teufel. 2010. *The structure of scientific articles: Applications to Indexing and Summarization*. CSLI Publications.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 21–26.

Peter Vinkler. 1998. Comparative investigation of frequency and strength of motives toward referencing. the reference threshold model. *Scientometrics*, 43(1):107–127.

Xiaojun Wan and Fang Liu. 2014. Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology (JASIST)*, 65(9):1929–1938.

Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science*, 342(6154):127–132.

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

Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1247–1252.

Rui Yan, Congrui Huang, Jie Tang, Yan Zhang, and Xiaoming Li. 2012. To better stand on the shoulder of giants. In *Proceedings of the 12th Joint Conference on Digital Libraries (JCDL)*, pages 51–60.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 594–604.

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427.

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

A Conversion of Teufel (2010) Data

As a part of training the classifier, instances from Teufel (2010) are used to supplement rare classes. Their data uses the scheme of Teufel et al. (2006b), which similar to our scheme but has several fine-grained distinctions. We convert the instances from their dataset as follows:

Teufel et al. (2006b) classification	Our Label
Weak	Comparison or Contrast
CoCoGM	Comparison or Contrast
CoCo	Comparison or Contrast
CoCoR0	Comparison or Contrast
CoCoXY	Background
PBas	Extends
PUse	Uses
PModi	Extends
PMot	Motivation
PSim	Comparison or Contrast
PSup	Comparison or Contrast
Neut	Background
CoMetN	Comparison or Contrast
CoGoaN	Comparison or Contrast
CoMet	Comparison or Contrast
CoCoN	Comparison or Contrast
CoCoM	Comparison or Contrast
CoResN	Comparison or Contrast

Note that we omit instances whose converted class is BACKGROUND in order to reduce the effects of a large majority class and because instances of the FUTURE class are merged into BACKGROUND according to their scheme.