Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme

Charles Jochim Hinrich Schütze
Institute for Natural Language Processing,
University of Stuttgart
charles.jochim@ims.uni-stuttgart.de

ABSTRACT
Citations are a valuable resource for characterizing scientific publications that has already been used in applications such as summarization and information retrieval. These applications could be even better served by expanding citation information. We aim to achieve this by extracting and classifying citation information from the text, so that subsequent applications may make use of it. We make three contributions to the advancement of fine-grained citation classification. First, our work uses a standard classification scheme for citations that was developed independently of automatic classification and therefore is not bound to any particular citation application. Second, to address the lack of available annotated corpora and reproducible results for citation classification, we are making available a manually-annotated corpus as a benchmark for further citation classification research. Third, we introduce new features designed for citation classification and compare them experimentally with previously proposed citation features, showing that these new features improve classification accuracy.

KEYWORDS: citation classification, feature extraction.

Proceedings of COLING 2012: Technical Papers, pages 1343–1358,
COLING 2012, Mumbai, December 2012.

1 Introduction

Citations are a valuable resource for characterizing scientific publications and their links to each other. They have been exploited for a number of natural language processing (NLP) and information retrieval (IR) applications, including summarization (Qazvinian and Radev, 2008; Qazvinian et al., 2010)[CJPF][1], improved indexing and retrieval (Ritchie et al., 2006)[CJPF]; and building integrated research databases (Nanba et al., 2004)[CJPF]. Bibliometric measures that quantify the impact of publications (e.g., Moed, 2005)[CJPF] are also based on citations.

Most of this work does not differentiate between uses of citations, e.g., whether a citation is more or less important to the paper or whether the paper's authors support or refute the claims made in the cited work. However, recently a number of research groups have attempted to classify citations with respect to dimensions like importance and relation to cited work (Teufel et al., 2006b; Dong and Schäfer, 2011; Sugiyama et al., 2010; Abu-Jbara and Radev, 2012)[CEPF]. By adding such fine-grained information to individual citations, the various applications of citation analysis can be better served; e.g., citations that are foundational to a paper may constitute better summary sentences for the cited paper.

Thus, there are clear potential benefits to fine-grained citation analysis; and a number of case studies have been published that demonstrate this potential (Nanba et al., 2004; Teufel et al., 2006b)[CEPF]. However, fine-grained citation analysis is currently not widely used in applications that access and analyze the scientific literature. In this paper, we identify a number of potential reasons for this state of affairs and propose solutions.

The first problem with current fine-grained citation analysis is that prior work has tended to develop custom classification schemes for a particular application. This means that the development cycle for a citation classifier must be started from scratch for each new application. In contrast to this prior work, we base our work on a standard classification scheme for citations from information science, the classification scheme of Moravcsik and Murugesan (1975)[CERF] (henceforth MM). We believe it is important to use an annotation scheme that is not bound to automatic citation classification for one particular task such as IR or bibliographic measures. Instead, it should be expressive enough to handle citations across many tasks. The MM scheme comprises four different dimensions or *facets*, which allows us to annotate the quality of the cited work along with its relation to the citing work. This gives the classification flexibility, so that it can be used in different application scenarios; e.g., some facets of the citation are more relevant for IR in digital libraries, while others are more useful in automatic summarization.

The second reason that fine-grained citation analysis has not seen widespread adoption is that it remains a challenge to accurately and automatically classify citations according to a predefined classification scheme (Teufel et al., 2006b)[CEPF]. We address this problem by introducing several novel features designed specifically for use in citation classification. Some of these new features are needed to support the more flexible and generic MM facet classification scheme. In particular, we extract novel features that capture the relationship between the citing paper and the cited paper. Identifying this relationship helps in understanding what motivated an author to reference the cited work. We also investigate how different features perform across the four facets, and how other variables, like the size of the context from which we extract features, affect the classification. We go on to compare different feature sets used for citation

[1]The citation annotation, described later in Sections 2 and 3, has likewise been applied to the citations in this paper. The following abbreviations apply: **C**=*conceptual*; **O**=*operational*; **E**=*evolutionary*; **J**=*juxtapositional*; **R**=*organic*; **P**=*perfunctory*; **F**=*confirmative*; **N**=*negational*.

classification. In particular we compare different lexical, syntactic, and positional features. To our knowledge this is the most extensive investigation of the comparative utility of features for citation analysis to date.

The final barrier to widespread adoption of fine-grained citation analysis is the fact that progress in the field has been hampered by the lack of a standard annotated corpus. Although all of the previous work we cover has used corpora of NLP articles for citation analysis experiments, none has tried reusing an existing corpus or annotation scheme. This makes accurately comparing results impossible, which in turn makes it difficult to gauge the advancement of the state of the art. Authors have focused on developing new annotation schemes, but no work has gone into building resources that allow the research community to evaluate and compare different citation classification methods.

As we will show below, results are also difficult or impossible to reproduce because existing citation approaches have not been described in sufficient detail and resources created or used for the approach have not been published. To address the lack of reproducible experiments in citation classification, we are making available, in conjunction with this paper, the manually-annotated corpus and feature vectors that produce the results reported here.[2] We hope that this corpus can provide a benchmark for further advances in citation classification.

The rest of the paper is structured as follows. Sections 2 and 3 cover the details of our annotation scheme and corpus, followed by a detailed description of the features used for classification in Section 4. Section 5 presents the different classification experiments we conduct with a discussion of the results in Section 6. Section 7 discusses related work. Finally, we close with a summary and an outline of future work.

2 Annotation scheme

In selecting our fine-grained classification scheme, we focused on two criteria. The first criterion is that we should consult the field of research that has the most expertise and the longest research record in developing classification schemes for citations. This field is information science. We have chosen the scheme proposed by Moravcsik and Murugesan (MM) because it adequately represents scientific literature for a broad range of citation classification scenarios. Furthermore, it is a well-established annotation scheme that is widely cited and used inside and outside of the information science community.

The second criterion for selecting the scheme was that it should be flexible and adaptable for different citation use cases. The MM scheme achieves this in that it is composed of four *independent* or *orthogonal* facets. For each facet, it assigns a label from a set of two labels. The scheme can be summarized with the four questions they posed: (i) Is the reference conceptual or operational? (ii) Is the reference organic or perfunctory? (iii) Is the reference evolutionary or juxtapositional? (iv) Is the reference confirmative or negational?

The *conceptual* vs *operational* facet – **CONC-OP** – asks: "Is this an idea or a tool?," where examples of tools are MRI in brain imaging and part-of-speech (POS) taggers in NLP. The *organic* vs *perfunctory* facet – **ORG-PERF** – distinguishes those citations that form the underpinnings of the citing work from more cursory citations. The *evolutionary* vs *juxtapositional* facet – **EVOL-JUX** – highlights the relationship between the citing and cited papers. If the citing paper builds on the cited work, it is EVOL while it is JUX if it presents an alternative to the cited

[2] `http://www.ims.uni-stuttgart.de/~jochimcs/citation-classification`.

work. Finally, **CONF-NEG**, the *confirmative* vs *negational* facet, captures the completeness and correctness of the cited work. A NEG citation usually is not derogatory, it may simply say that the cited work is weaker than the citing work or is otherwise missing some critical point. These distinctions are covered in more detail in the annotation guidelines.[2]

These four facets can be thought of as orthogonal dimensions along which citations can vary. This is the basis for flexible and adaptable citation analysis; e.g., a facet that is not relevant for a particular application can simply be omitted. If interactions between two facets are important for another application, they are made available by the citation classifier without complicating the model or its training.

Although there are now four facets to annotate for each citation instead of a single label, the annotation task is not more difficult. Making a binary decision is easier than trying to pick a label from ten possibilities with subtle differences between some of them. Yet, with the combination of different facets we still can achieve a finer-grained label.

It is also important to note that this classification has no undefined class. Several previous annotation schemes have a default label, *neutral* or *other*, that is assigned to a citation when no other classes can be. In the work we have seen that uses such annotation schemes, more than half of the citation instances are assigned this undefined label. In these cases, summarization or IR systems that want to make use of citation information obtain no useful information from the citation classifier for more than half of citations.

3 Corpus

Our corpus, like corpora from some previous studies (Athar, 2011; Dong and Schäfer, 2011)[CEPF], is taken from NLP literature. Specifically, we have taken the 2004 ACL proceedings from the ACL Anthology Reference Corpus (ARC) (Bird et al., 2008)[OEPF]. NLP literature was chosen because our annotators (NLP students) are more familiar with this data and can make more informed decisions when annotating the citations.

Some statistics on the number of documents and citations in the corpus can be found in Table 1. Each citation in the corpus has been independently annotated by at least two of six annotators. Gold labels are chosen by a simple majority vote and in the case of ties the votes of more experienced annotators are weighted higher. The annotators were given guidelines to help ensure consistent annotation. We built a browser-based annotation tool that displays the full text of the paper, so that the annotators can look at the wider context of the citation when necessary. In many cases the context necessary for annotation is only one sentence, but it will often span sentences or fill a paragraph.

section	docs	citations
main ACL	57	1668
student	6	101
poster/demo	21	239
total	84	2008

CONC	OP	EVOL	JUX
1792	216	1804	204
ORG	**PERF**	**CONF**	**NEG**
203	1805	1836	172

Table 1: ACL 2004 corpus (left) and summary of annotated citations (right).

As mentioned in Section 2, no facets were left undefined. This reduces the classification to only two classes and avoids a neutral class. For our purposes it is reasonable to avoid having a neutral class; e.g., a citation that is not explicitly CONF should still be implicitly considered CONF because including the citation is still an endorsement of the cited work.

Fleiss's κ values of the annotation are .42 (CONC-OP), .45 (EVOL-JUX), .18 (ORG-PERF) and .41 (CONF-NEG). These numbers indicate that the difficulty of the annotation task varies for the different facets, with ORG-PERF being most difficult.[3] Due to the highly skewed distribution, κ suffers from *prevalence* (Eugenio and Glass, 2004)[CEPF]; yet three of the facets still have *moderate* agreement (according to Landis and Koch (1977)[CEPF]), and ORG-PERF has *slight* agreement. We feel that the observed agreement[4] is high enough that we can rely on the gold labels for evaluation.

We are releasing the corpus along with this paper.[2] To the best of our knowledge this corpus is the first to be annotated by individuals other than the study's authors. It is important to have independent annotators to limit any bias in the gold-standard annotation. One consequence of this is that our inter-annotator agreement scores are lower than those previously published as the previous annotation came from the developers of the respective annotation schemes and from the authors reporting on the classification experiments using them.

4 Description of features

Our goal is to accurately classify citations according to MM, the annotation scheme described in Section 2. We make the assumption that the necessary clues for correctly labeling citations, both manually and automatically, can be found in the context of the citation, i.e., the running text surrounding the citation. If we are able to extract the right clues from the citation context we can accurately label the citation's use.

Because there is not yet a standard corpus for the task of automatic citation classification, the results from previous work are difficult to compare. Previous studies have used different corpora, different annotation schemes, different feature sets, and different classifiers. In an effort to borrow from – and eventually compare ourselves to – previous work, we investigate some of the features used previously and introduce our own. The reader may want to refer to the overview of features in Table 2 as we describe the features in what follows.

Lexical features. Much of the earlier work on automatic citation classification (Dong and Schäfer, 2011; Nanba and Okumura, 1999; Teufel et al., 2006b)[CEPF] relied on cue words and phrases ($\texttt{cues}_k$). These were often implemented as follows. For a class (e.g., Dong and Schäfer's *idea* class), a list of of cues (e.g., the word "following") are defined that indicate that class. Finally, a Boolean feature (e.g., $\texttt{cues}_{idea}$) is set to true if any word from the list is in the citing context. This results in k Boolean features where k is often the number of classification labels (although it can be greater, see Dong and Schäfer (2011)[CEPF]).

Different length n-grams were later used by Athar (2011)[CJPF] with results indicating that combined unigram, bigram, and trigram features (`1+2+3-gram`) performed better than unigrams (`1-gram`) and unigrams plus bigrams (`1+2-gram`).

We use only unigrams because they perform at least as well as using unigrams, bigrams and trigrams in our experiments, without introducing a much larger, sparsely-populated feature set. Unigrams should also be quite robust and perform reasonably well across the four facets.

Word-level linguistic features. Part-of-speech (POS) tags of the words in the citation sentence were used as features by Athar (2011)[CJPF] (`POS` and `1-gram+POS`). Select linguistic features

[3] MM's definition was "is the reference truly needed for the understanding of the referring paper," so the annotation hinges on the understanding of the individual annotator, resulting in higher disagreement.

[4] Agreements were: .86 (CONC-OP), .88 (EVOL-JUX), .72 (ORG-PERF), .91 (CONF-NEG)

feature class	name	source	type or *example value*	description
lexical feats.	cues$_k$	NO99, TST06, DS11	Boolean	k Boolean features: one for each group of cue words/phrases
	1-gram	Ath11, own	*hard*	unigrams
	1+2-gram	Ath11	*hard language*	unigrams & bigrams
	1+2+3-gram	Ath11	*hard language like*	unigrams, bigrams, & trigrams
word-level linguistic feats.	POS	Ath11	*NN, JJ, IN*	POS tags
	1-gram+POS	Ath11	*quality+NN, new+JJ*	POS tag-word conjunctions
	tense	TST06	*present, past*	verbal tense
	voice	TST06	*active, passive*	verbal voice
	modal	TST06	*can, may*	modal verb (if any)
	has-modal	own	Boolean	sentence has modal verb
	root	own	*have, present*	dependency root node
	main-verb	own	*present, use*	main verb
	has-1stPRP	own	Boolean	first person POS
	has-3rdPRP	own	Boolean	third person POS
	comp/sup	own	*more, better*	comparative/superlative POS
	but	own	Boolean	has "but"
	has-cf	own	Boolean	has "cf."
ling. structure feats.	dep-rel	Ath11	*pobj:to:information*	Stanford typed dependencies (de Marneffe et al., 2006)
	POS-pattern$_k$	DS11	Boolean	k Boolean features: one for each POS tag pattern
	is-constituent	own	Boolean	citation is a constituent
	self-comp	own	Boolean	author linked to comparative
	other-comp	own	Boolean	citation linked to comparative
	other-contrast	own	Boolean	citation is in contrastive clause
	self-good	own	Boolean	author linked to positive sentiment
location feats.	section	DS11	*Introduction, Method*	1 of 6 possible section headings
	paper-loc	TST06	unknown	citation position in paper
	paragraph-loc	TST06	unknown	citation position in paragraph
	section-loc	TST06	unknown	citation position in section
	sentence-loc	own	*beginning, middle, end*	location in the first quarter, middle half (25%-75%), and last quarter
frequency feats.	popularity	DS11	Integer	citations in the same sentence
	density	DS11	Integer	citations in the same context (sentence and its neighbors)
	avgDensity	DS11	Real	average density of neighboring sentences
sent. feats.	scilex	Ath11	unknown	scientific polarity lexicon
	cpol	Ath11	unknown	general polarity lexicon
	positive-words	own	*best, advantage*	general positive lexicon
	negative-words	own	*problem, against*	general negative lexicon
other feats.	self-cite	TST06	Boolean	citation to own work
	has-resource	own	Boolean	resource entity found with NER
	has-tool	own	Boolean	tool entity found with NER

Table 2: Feature list (grouped by feature class). NO99=Nanba and Okumura (1999); TST06=Teufel et al. (2006b); Ath11=Athar (2011); DS11=Dong and Schäfer (2011). "unknown" = exact definition of the feature (e.g., Boolean or Real) is unknown. Examples of possible feature values are given in italics where appropriate.

related only to the main verb were shown to be effective by Teufel et al. (2006b)[CEPF]; e.g., tense (`tense`), voice (`voice`), and modality (`modal`).

We also include modality in our feature set (`has-modal`) along with separate features for the main verb (`main-verb`) and the root (`root`) as determined by the MATE dependency parser (Bohnet, 2010)[OEPF]! We do not include POS as features per se, but some features are triggered by the occurrence of selected POS: 1st and 3rd person pronouns (`has-1stPRP`, `has-3rdPRP`); and comparatives and superlatives (`comp/sup`). Comparatives and superlatives can help distinguish CONF from NEG. Pronouns on the other hand may be useful in classifying EVOL-JUX, e.g., first person pronouns are used when clarifying the differences between proposed and cited approaches. We add two other features for the contrastive conjunction "but" (`but`) and the abbreviation "cf." (`has-cf`). In our analysis of citations we looked at the role of contrastive conjunctions in citation sentences and found these simple features to be useful.

Linguistic structure features. Dependency relations (`dep-rel`) were used as features and showed a marked improvement over the baseline by Athar (2011)[CEPN]! Dong and Schäfer (2011)[CEPF] used seven regular expression patterns of POS tags ($\texttt{POS-pattern}_k$) to capture syntactic information (e.g., ".*(VHP|VHZ) VV.*"); then $k = 7$ Boolean features marked the presence (or absence) of these patterns.

We add other new features related to the linguistic structure of the citation sentence. For `is-constituent`, the citation is labeled as a constituent if the authors appear outside of the parentheses with only the date in parentheses, e.g., "Gusfield (1997) showed that . . . ", or if the citation acts as a placeholder for the cited work following a preposition, e.g., ". . . following the experiments in (Kaplan et al., 2004)". These cases are distinguished from citations like: ". . . are two popular examples of kernel methods (Fukunaga, 1990; Cortes and Vapnik, 1995)". We are relying here on a certain style of writing and citation format, like that found in ACL proceedings. We expect this feature to help for ORG-PERF as organic citations are more likely to show up as constituents in citation sentences.

The personal pronoun and comparative features mentioned above (`has-1stPRP`, `has-3rdPRP`, and `comp/sup`) are useful features, but we would like to extract a more specific feature that links them. We want features that indicate that the citing work is better than the cited work. To obtain these features we parse the sentence and extract relations from the parse tree. For the author/comparative relation, we first find the comparative in the sentence and traverse the tree to find the subject of the phrase that contains that comparative. If the subject refers to the author of the paper (e.g., with a first person pronoun), we set the `self-comp` feature to true.

We also found that JUX citations are often set apart using contrastive conjunctions, e.g., *while* or *despite*. We again traverse the parse tree to extract the relationship between contrastive conjunctions and the citation (`other-contrast`), where the citation or cited authors show up in the dependent clause governed by the contrastive conjunction. The feature is set to true if the citation is found among the descendants of the contrastive conjunction.

Location features. The section of the paper in which the citation is located (`section`) was used as a feature by Dong and Schäfer (2011)[CEPF]! Teufel et al. (2006b)[CEPF] also included location features at different granularities: within the paper (`paper-loc`), within the paragraph (`paragraph-loc`), and within the section (`section-loc`).

We include a different location feature approximating where the citation is found in the sentence (`sentence-loc`): beginning, middle, or end. This feature is motivated by the fact that citations

at the end of the sentence are predominantly PERF.

Frequency features. Dong and Schäfer (2011)[CEPF] used the number of citations in a single sentence (`popularity`) and in the citation sentence plus its neighboring sentences (`density`) as features. They also included a third feature for the average density of neighboring sentences (`avgDensity`).

Sentiment features. Athar (2011)[CEPF] included two different polarity lexicons. One is hand-crafted and specific to the scientific domain (`scilex`). The other is the large general purpose polarity lexicon from Wilson et al. (2005)[OEPF] (`cpol`). He also tried features (`neg`) that account for negation. This was done by appending "_neg" to the end of the 15 lexical items that follow any negation term.

We were not able to obtain the scientific polarity lexicon, but use the polarity lexicon from Wilson et al. (2005)[OEPF] to extract sentiment features. Our polarity features are represented as a bag of words (BOW) where the citation context words present in the polarity lexicon are added to the BOW features `positive-words` or `negative-words` according to their polarity. Although CONF-NEG is not strictly a matter of sentiment, we still apply this feature hoping for improvements on this facet.

Self-reference feature. Teufel et al. (2006b)[CEPF] used a feature, `self-cite`, that indicates if one of the citing authors also (co-)authored the cited work. This feature is unique in that it is the only feature based on the reference and not the individual citation and therefore not taken from the context in which it is found.

NER features. Using lexical features alone, there are a number of words that help indicate OP (*operational*) citations in NLP, e.g., "parser", "tagger", "corpus". We decide to take this a step further and train a named-entity recognition (NER) system to identify NLP named entities. We identify two types of NLP named entities: corpora and tools. First, we create a gazetteer of NLP tools and corpora from an online list of these resources.[5] Next, we tag a portion of our corpus using the gazetteer list to label any occurrence of the words in the list and then manually check those labeled instances to be sure they are correctly labeled. In this way we can expediently create training data, with an emphasis on precision over recall. Finally, we train the SuperSenseTagger (Ciaramita and Altun, 2006)[OEPF] on this annotated portion, and tag the remaining part of the corpus. NER is not central to our task, so we did no direct evaluation of it; we looked only to see if it might lead to improvements in our classification. We include two features, `has-resource` and `has-tool`, for the two types of entities.

The NER features we extract are related only to the NLP domain. However, this approach for acquiring named entities is not domain dependent and can be used to develop a reasonably efficient NER system using lists of tools or resources from any domain.

5 Experiments

In this section we will outline our classification experiments and then discuss the results in Section 6. We use the term *feature set* to describe a collection of features used by us or in previous studies; we use the term *feature class* to describe a collection of similar features as they are organized in Section 4 and in Table 2.

Setup. All our experiments were conducted on the corpus described in Section 3. We trained the Stanford MaxEnt classifier (Manning and Klein, 2003)[OEPF] for each of the four facets

[5] `http://nlp.stanford.edu/links/statnlp.html`

in a 5-fold cross validation setup with default settings except that we set the regularization parameter $\sigma = 10$ based on previous experiments.

Feature set comparison. In our first set of experiments we test our own feature set and the feature sets described in previous studies. Each of these feature sets is a subset of the features described in Section 4 and is identified below by some of its more distinguishing features; e.g., **NgramDep** refers to the feature set that mainly uses n-grams and dependencies.

CueVerbLoc. This feature set is intended to mimic (Teufel et al., 2006b)[CEPF] to the extent this is possible. It includes cue phrase features ($\texttt{cues}_k$), the verbal features `tense`, `voice`, and `modal` as well as `paper-loc`. The cue phrases used in (Teufel et al., 2006b)[CJPF] are not available so we applied automatic feature selection using mutual information (MI) (Manning et al., 2008)[CEPF] to select the most informative unigrams, bigrams, and trigrams for each class label. We borrow from the manual feature selection in (Teufel et al., 2006b)[CEPF] by assigning cue phrases to each of the labels (8 in our case – Teufel et al. used 12) and limiting the number of cue words to 75 per label. Some examples for OP cues are *wordnet*, and *parser*.

NgramDep. This feature set corresponds to (Athar, 2011)[CEPF]. It includes lexical features: unigrams, bigrams, and trigrams (`1+2+3-grams`) and the `dep-rel` features. Athar (2011)[CJPF] tested other features, but we have only reimplemented those that improved results.

CueFreqPOS. This feature set is based on (Dong and Schäfer, 2011)[CEPF]. It includes a list of cue words ($\texttt{cues}_k$), then the frequency features `popularity`, `density`, `avgDensity`, and the syntactic feature $\texttt{POS-pattern}_k$.

PREV. This feature set combines all features previously used for citation classification into one feature set (i.e., CueVerbLoc + NgramDep + CueFreqPOS).

OWN. The feature set OWN includes all the features we have introduced in our work – those marked "own" in Table 2. Some features were designed to help one facet or another, but we use them all together here for all facets.

We note here that by reimplementing features from previous work we claim only to extract the same or similar information as the original authors. Due to sometimes major differences in the corpus, annotation scheme, and classifier used, we are not able to reproduce the same conditions that led to previous results. We are instead more interested in the types of features that seem to perform best on our dataset with our annotation scheme.

Citation context size. The tests just described are run with a fixed context size of one sentence. It is not clear how much context is best for feature extraction, so in another set of experiments we fix the feature set and test the features extracted from different sized context windows. In previous work, different sized context windows were used by different studies, e.g., Athar (2011)[CEPF] used only the sentence containing the citation while Dong and Schäfer (2011)[CEPF] used up to three sentences. Kaplan et al. (2009)[CEPF] and Abu-Jbara and Radev (2012)[CEPF] have illustrated the difficulties in delineating the exact boundary for each individual citation context, while Athar and Teufel (2012)[CEPF] tried different fixed context sizes for citation classification. We follow this general idea and test context lengths of 1, 2, and 3 sentences.

Feature class comparison. In addition to comparing our own feature set with those from previous work, we also want to investigate what feature classes assist most in the classification. We perform this analysis by examining the impact of the seven feature classes described in Section 4. More specifically, we compare the results of their individual performance using *only features in the feature class* (Table 4, top), and their ablation from the entire feature set using *all*

features except those in the feature class (Table 4, center). Finally, we extend the ablation study, *successively removing all feature classes* in order of importance (i.e., by their contribution to F_1 score) (Table 4, bottom).

6 Results and Discussion

Feature set results. The results for the different feature sets when using one sentence of context are found in Table 3. All of the F_1 results presented in this paper are macro-averaged F_1. We have included two baseline experiments. We use a majority baseline (BL) that labels each citation with the label occurring most often in the corpus, e.g., for CONC-OP, all citations are labeled CONC. We also include results for unigram, bigram, and trigram features (Ngram), which is the baseline used by Athar (2011)[CJPN]. The results in Table 3 show that our feature combination outperforms both baselines and all reimplemented feature sets for all four facets. With two exceptions (Ngram for EVOL-JUX and PREV for ORG-PERF), these results are significant.[6]

The greatest improvement over the baseline is with the OWN features for CONC-OP. Several of the other feature sets also do better on CONC-OP than BL, but OWN is still significantly better than PREV, the combination of all other feature sets. Simple BOW features along with our new features (e.g., `has-resource` and `has-tool`) increase F_1 by 7 points over PREV. As an example, in a sentence citing "The Penn TreeBank (Marcus et al., 1993)," the citation is incorrectly classified using PREV. The NER tool recognizes Penn TreeBank as a corpus, which results in the OWN feature `has-resource` to be set to true and a correct classification of the citation as OP.

EVOL-JUX proves to be more difficult than CONC-OP with either no or very small improvements over BL for all feature sets except for Ngram and OWN. The BOW features from our OWN feature set are responsible for most of the improvement of F_1 from 47.3 to 52.9. BOW features contribute to the improvement with OWN features for all four facets.

OWN features improve F_1 by 10.7 (from 47.3 to 58.0) over the BL for ORG-PERF, and are also better by 3.2 (54.8 vs 58.0) than PREV. Some features that contribute to the better results are `root` and `main-verb` with values such as "describe" and "present"; these appear to be useful in identifying ORG citations. In this facet, the feature set CueFreqPOS sees its most significant improvement over BL. This is due in a large part to the frequency features that are not found in other feature sets.

Finally, CONF-NEG is the most difficult facet. All feature sets except our own performed only as well as or even worse than BL. OWN features improve F_1 by 3.3 (from 47.8 to 51.1), which is due in part to the location feature that finds citations in the middle of sentences to be CONF, while NEG citations are more likely to come at the beginning.

To get an idea of a possible upper bound for this task, we include a *human classifier* ("Human" in Table 3): we take the annotation from the most experienced annotator and consider it as classification output. CONC-OP is the "easiest" facet for the human classifier to label, similar to automatic classification. However, the most difficult facet for automatic classification, CONF-NEG, appears to be straightforward for the human classifier. This is consistent with the high observed agreement for CONF-NEG (.91, footnote 4).

Context size results. For OWN, we tested three different context sizes c: $c \in \{1, 2, 3\}$ sentences. We found that $c = 1$ is best for CONC-OP (significant) and ORG-PERF; and $c > 1$ is better for

[6] $p < .05$. All significance tests in this paper use the approximate randomization test (Noreen, 1989)[CEPF].

	CONC-OP	EVOL-JUX	ORG-PERF	CONF-NEG
baseline (BL)	*47.2	*47.3	*47.3	*47.8
Ngram	*53.2	50.7	*51.3	*47.8
CueFreqPOS	*48.4	*49.4	*54.1	*47.7
NgramDep	*53.3	*47.3	*50.5	*47.8
CueVerbLoc	*51.1	*47.3	*47.3	*47.8
PREV	*61.2	*48.5	54.8	*47.5
OWN	68.2	52.9	58.0	51.1
Human	94.7	91.1	91.7	93.5

Table 3: F_1 for different feature sets. Marked with $*$: significantly worse than OWN ($p < .05$). Underlined: best performing feature set per facet.

CONF-NEG (significant) and EVOL-JUX. These results suggest that context size is an important factor, but one that does not have a uniform effect on the four facets. The online appendix describes these experiments in more detail.[2]

Feature class results. In the discussion of the feature class results we will refer to the line numbers in Table 4. The table presents F_1 results using only a single feature class (lines 1–7); F_1 using all features ("All") and F_1 using all features except the listed feature class (lines 8–14); and finally, extended ablation results where a feature class is successively removed from "All" (seven classes) until one feature class remains (lines 15–21). Our goal is to get a better idea of which feature classes are informative for a given facet.

CONC-OP. LEXICAL features appear to be the most important for this facet. Alone they do well against the baseline (61.6 vs 47.2, line 1) and when removed from the entire feature set F_1 drops more than for any other feature class (from 64.5 to 58.2, line 8). Both of these Δ's are significant. The feature class NER has the second highest F_1 (54.1, line 7) when used alone, which makes sense as it was designed for this facet. Removing NER features hurts F_1 (down to 64.0, line 14), but not significantly. Using only WORD-LEVEL or STRUCTURE features also leads to significant improvement: increases of 4.8 (line 2) and 4.3 (line 3). After that, SENTIMENT features improve F_1 but not significantly (line 6), while the LOCATION and FREQUENCY features show no difference from the BL (lines 4–5). The ablation results show that after the significant contributions of the LEXICAL features, the removal of other feature classes does not affect the results much: Removing STRUCTURE, LOCATION, and SENTIMENT features actually increases F_1 (lines 10, 11, 13), and the ablation of WORD-LEVEL, FREQUENCY, and NER features shows no significant change (lines 9, 12, 14).

EVOL-JUX. For this facet, three of the seven feature classes, LOCATION, FREQUENCY, and NER, lead to no change from the baseline when run alone (lines 4, 5, 7). Another three feature classes, LEXICAL, WORD-LEVEL, and SENTIMENT, significantly improve over BL (lines 1, 2, 6). Conversely, the FREQUENCY features, with no improvement alone, help improve results of the entire feature set; when those features are removed, F_1 drops by 2.2 (from 53.4 to 51.2, line 12). Also, the SENTIMENT features, which do well against the baseline (line 6), hurt F_1 when added to the full feature set (decrease by -0.7, line 13).

ORG-PERF. Individually, the feature classes LEXICAL, WORD-LEVEL, and STRUCTURE all had significant improvements (lines 1–3). The other four classes do not help for this facet (lines 4–7). However, in the ablation results, omitting these feature classes also *increases* F_1 (lines 8–10). Only removing LOCATION significantly decreases F_1 (line 11). This result indicates that several of the feature classes are correlated for classifying this facet. They contain useful information for the task (as indicated by good performance when used individually), but mutual correlation has

the effect of bad generalization when all of them are used together. The results show that this type of analysis (which has not been performed before for citation classification) is important to understand how features impact performance and what steps are needed to achieve better performance.

		CONC-OP		EVOL-JUX		ORG-PERF		CONF-NEG	
	BL	47.2		47.3		47.3		47.8	
		F_1	Δ BL	F_1	Δ BL	F_1	Δ BL	F_1	Δ BL
1	LEXICAL	†61.6	†14.4	†52.7	†5.4	†56.1	†8.8	47.7	0.0
2	WORD-LEVEL	†52.0	†4.8	†52.4	†5.0	†51.6	†4.2	49.7	2.0
3	STRUCTURE	†51.5	†4.3	48.8	1.5	†52.0	†4.7	47.8	0.0
4	LOCATION	47.2	0.0	47.3	0.0	47.3	0.0	47.8	0.0
5	FREQUENCY	47.2	0.0	47.3	0.0	47.3	0.0	47.8	0.0
6	SENTIMENT	48.0	0.9	†52.7	†5.3	47.2	-0.1	†49.9	†2.1
7	NER	†54.1	†7.0	47.3	0.0	47.3	0.0	47.8	0.0

		CONC-OP		EVOL-JUX		ORG-PERF		CONF-NEG	
	All	64.5		53.4		59.2		48.9	
		F_1	Δ All	F_1	Δ All	F_1	Δ All	F_1	Δ All
8	LEXICAL	*58.2	*6.2	53.3	0.1	60.2	-1.0	49.5	-0.6
9	WORD-LEVEL	64.0	0.4	53.6	-0.3	59.3	-0.1	48.9	0.1
10	STRUCTURE	66.7	-2.2	53.1	0.3	59.5	-0.3	*48.8	*0.1
11	LOCATION	65.0	-0.5	53.2	0.1	*55.8	*3.4	49.6	-0.6
12	FREQUENCY	64.4	0.1	51.2	2.2	58.3	0.9	49.8	-0.9
13	SENTIMENT	65.0	-0.5	54.1	-0.7	59.2	0.0	49.0	0.0
14	NER	64.0	0.4	53.7	-0.3	58.8	0.4	49.4	-0.5

	CONC-OP		EVOL-JUX		ORG-PERF		CONF-NEG	
	All	64.5	All	53.4	All	59.2	All	48.9
15	LEXICAL	*58.2	FREQUENCY	51.2	LOCATION	*55.8	STRUCTURE	*48.8
16	NER	*53.6	STRUCTURE	*50.3	LEXICAL	*55.4	WORD-LEVEL	48.2
17	STRUCTURE	*50.3	LEXICAL	50.6	STRUCTURE	*53.0	LOCATION	47.4
18	WORD-LEVEL	*47.9	NER	50.7	WORD-LEVEL	*48.6	NER	47.4
19	SENTIMENT	*47.2	LOCATION	52.7	SENTIMENT	*47.3	SENTIMENT	47.4
20	FREQUENCY	*47.2	SENTIMENT	52.4	FREQUENCY	*47.3	FREQUENCY	47.7
21	LOCATION	*47.2	WORD-LEVEL	*47.3	NER	*47.3	LEXICAL	47.8

Table 4: **Top.** Results when a single feature class is used. **Middle.** Ablation results: F_1 and decrease in F_1 when each feature class is ablated; i.e., each result shown is a classification result using six feature classes. **Bottom.** Extended ablation results: Left columns indicate the feature class removed. Marked with †: significantly better than BL ($p < .05$); marked with $*$: significantly lower than All ($p < .05$). Underlined: best performing feature class per facet (largest Δ).

CONF-NEG. Only SENTIMENT (line 6) and WORD-LEVEL (line 2) improve over BL and the remaining five feature classes do only as well as BL. Removing four of the seven feature classes actually seems to improve F_1 (lines 8, 11, 12, 14), with F_1 only increasing by adding WORD-LEVEL or STRUCTURE features (lines 9–10).[7] In fact, it seems that including the feature classes LEXICAL, STRUCTURE, LOCATION, FREQUENCY, and NER might only be detrimental for this facet, as F_1 using only SENTIMENT features is 49.9 (line 6) compared to using all features at 48.9 ("All").

To further analyze the relative importance of a feature class for a facet we extend the ablation results by successively removing that feature class whose removal results in the lowest F_1, among the possible ablations, until all have been removed (Table 4, lines 15–21). E.g., in CONC-OP we start with all features ($F_1 = 64.5$) and calculate F_1 after removing each of the feature classes individually. In this case, removing LEXICAL leads to the largest drop in F_1, from

[7] Lines 9–10 have different F_1 (48.9 vs 48.8) but the same $\Delta = 0.1$ due to rounding.

64.5 to 58.2 (line 15). In the next iteration, we again compare F_1 after removing each of the six remaining feature classes. Removing NER features results in the lowest F_1 (now 53.6, line 16), and we proceed by removing one of the five remaining feature classes, etc. These results support what was discussed for the top and middle portions of Table 4, but present it as a list of the feature classes in descending order of importance. This table helps us to compare different facets; we can easily see that LEXICAL and NER features are important for CONC-OP, while LOCATION features are not. Compare this to CONF-NEG where LEXICAL and NER features are not important and WORD-LEVEL is higher in the list. Note also, that F_1 does not always decrease (e.g., removing LEXICAL for EVOL-JUX). Some combinations of subsets of features will perform better than the previous superset. In this case, we see that after having removed STRUCTURE, removing any other feature class can only improve results.

The results in this section give us some valuable insight into how to design features for citation classification. First, we consider the first three feature classes, LEXICAL, WORD-LEVEL, and STRUCTURE. All three contain quite general text classification features, and consequently are quite robust and informative across the four facets of citations that we consider. WORD-LEVEL seems to be the most robust across all four facets, while LEXICAL has the largest Δ BL values for three of the four facets (i.e., CONC-OP, EVOL-JUX, and ORG-PERF). The last four feature classes – LOCATION, FREQUENCY, SENTIMENT, NER – represent different citation features which seem to impact certain citation facets. NER was designed particularly for CONC-OP and does in fact contribute most to that facet; LOCATION helps only ORG-PERF (i.e., the position of the citation indicates its importance) where it contributes significantly to a combination of features; similarly FREQUENCY contributes significantly to a combination of features for EVOL-JUX; and finally, SENTIMENT is important for EVOL-JUX and CONF-NEG, as expected. There is no single feature class that is the most important for all facets, which lends credence to the claim that these facets capture different properties of citations. We conclude that our multi-faceted scheme benefits from a diverse feature set and that although general, easily-extractable features help classification more consistently, the extraction of more specific features is important for improvements on certain classification tasks.

7 Related Work

Information scientists started labeling and studying citations long before automatic text classification became a reality. Garfield (1964)[CEPF] originally introduced 15 different motivations for why an author might cite a paper; Weinstock (1971)[CEPF] then revisited this classification as he explored the emergence of citation indexes. Several studies following Weinstock also aimed to characterize the *function* of citations (as opposed to the motivation). One example is the MM scheme we adopt here. Chubin and Moitra (1975)[CEPF] attempted to simplify and flatten MM using six categories. Spiegel-Rösing (1977)[CEPF] produces another classification scheme with 13 categories that she uses to evaluate one journal's scholarly contributions. Further comparison of previous citation studies can be found in (Liu, 1993)[CEPF] and more recently in (Bornmann and Daniel, 2008)[CEPF]. We note that several other studies (Cano, 1989; McCain and Turner, 1989)[CEPF] have also reused or refined MM in some way, which reinforces our choice. As stated earlier in Section 2, we feel that the multi-faceted composition of MM provides us with a more flexible annotation scheme and a powerful one that can easily represent the quality of a citation as well as its relation to the citing author.

These early annotation schemes were manually applied to a limited amount of scientific literature and did not consider automatic application on large amounts of text. One early

application of automatic citation classification (Nanba and Okumura, 1999)[CEPF] used an annotation scheme with only three classes (Basis, Compare, Other) that are reportedly based on the 15 classes from Weinstock (1971)[CEPF]. Teufel et al. (2006a)[CEPF] introduce a much more complete annotation scheme with 12 classes designed for IR. They thoroughly motivate and analyze their annotation scheme and report inter-annotator agreement of κ=.72. More recently, sentiment analysis has been applied to citations. Athar (2011)[CEPF] classifies citations as *positive*, *negative*, and *objective*, and finds marked improvement in classification using dependency relation features. Athar and Teufel (2012)[CEPF] extend this work and consider context windows of different widths. For each of these three studies the largest class is the one with the least informative label: Nanba and Okumura's *Other* is 52% of citations; Teufel et al.'s *Neutral* is 63%; and Athar's *objective* is 86%. This means that an application receives little information about a majority of citations. In contrast, our annotation scheme does not have a neutral label and always assigns a multi-faceted label that will contain some useful information as no facet can be left undefined.

Dong and Schäfer (2011)[CEPF] conducted a classification study using their own classification scheme with four labels relating to the function of the MM *organic/perfunctory* facet. In addition to adding new syntactic features (`POS-pattern`$_k$, see above), they tested ensemble-style self-training to overcome the problem of limited annotated data. Their paper also included a new dataset with annotated citing sentences. It is important to use previously-tested, publicly-available data, however, their dataset does not contain the full corpus from which they extracted features. Due to this restriction we cannot extract many of the features that they use (e.g., features in the LOCATION and FREQUENCY classes). The annotation in their dataset is also attached to the sentence and not individual citations. This makes it impossible to classify individual citations and prevents us from using the citation-specific features that we have developed (OWN features in STRUCTURE class, e.g., `is-constituent`). We have conducted experiments on the Dong and Schäfer dataset and include those experiments in the online appendix.[2] We believe that annotating and classifying citing sentences (as opposed to citations) is not specific enough for tasks like IR and bibliometrics. Thus, it is essential that we have a citation-annotated corpus for accurate classification.

As we have argued above, the motivation for our work is to provide a generic classification scheme that is established and accepted in information science in the hope that it can be used for a wide range of applications.

Conclusion

In this paper, we address the task of citation classification for applications that access and analyze the scientific literature. Our work uses MM, a standard classification scheme for citations that was developed independently of automatic classification and therefore is not bound to any particular citation application. We introduce new features designed for citation classification and show that they improve performance as measured by F_1. To address the lack of available annotated corpora and reproducible results for citation classification, we are publishing, along with this paper, a manually-annotated corpus as a benchmark for further citation classification research. In future work, we want to further extend the feature set to improve classification and show the benefits of our system for applications like bibliometrics.

Acknowledgments. We thank DFG for funding this work (SPP 1335 *Scalable Visual Analytics*) and Christian Scheible, Wiltrud Kessler, Alex Fraser and the anonymous reviewers for their contributions to the paper.

References

Abu-Jbara, A. and Radev, D. R. (2012). Reference scope identification in citing sentences. In *Proceedings of HLT-NAACL*, pages 80–90.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL Student Session*, pages 81–87.

Athar, A. and Teufel, S. (2012). Context-enhanced citation sentiment detection. In *Proceedings of HLT-NAACL*, pages 597–601.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of LREC*, pages 1755–1759.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING*, pages 89–97.

Bornmann, L. and Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, 40:284–290.

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

Ciaramita, M. and Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP*, pages 594–602.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, pages 449–454.

Dong, C. and Schäfer, U. (2011). Ensemble-style self-training on citation classification. In *Proceedings of IJCNLP*, pages 623–631.

Eugenio, B. D. and Glass, M. (2004). The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

Garfield, E. (1964). Can citation indexing be automated? In *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings*, pages 189–192.

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 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 88–95.

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

Liu, M. (1993). Progress in documentation the complexities of citation practice: A review of citation studies. *Journal of Documentation*, 49(4):370–408.

Manning, C. D. and Klein, D. (2003). Optimization, maxent models, and conditional estimation without magic. In *Proceedings of HLT-NAACL*, pages 8–8.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York.

McCain, K. W. and Turner, K. (1989). Citation content analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, 17(1–2):127–163.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Springer.

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

Nanba, H., Abekawa, T., Okumura, M., and Saito, S. (2004). Bilingual PRESRI - integration of multiple research paper databases. In *Proceedings of RIAO*, pages 195–211.

Nanba, H. and Okumura, M. (1999). Towards multi-paper summarization using reference information. In *Proceedings of IJCAI*, pages 926–931.

Noreen, E. (1989). *Computer-intensive methods for testing hypotheses: An introduction*. Wiley.

Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of COLING*, pages 689–696.

Qazvinian, V., Radev, D. R., and Özgür, A. (2010). Citation summarization through keyphrase extraction. In *Proceedings of COLING*, pages 895–903.

Ritchie, A., Teufel, S., and Robertson, S. (2006). How to find better index terms through citations. In *Proceedings of Workshop on How Can Computational Linguistics Improve Information Retrieval?*, pages 25–32.

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

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

Teufel, S., Siddharthan, A., and Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of SIGdial Workshop on Discourse and Dialogue*, pages 80–87.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006b). Automatic classification of citation function. In *Proceedings of EMNLP*, pages 103–110.

Weinstock, M. (1971). *Encyclopedia of Library and Information Science*, volume 5, chapter Citation indexes. Dekker, New York, NY.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*, pages 347–354.