# Purpose and Polarity of Citation: Towards NLP-based Bibliometrics

**Amjad Abu-Jbara**
Department of EECS
University of Michigan
Ann Arbor, MI, USA
amjbara@umich.edu

**Jefferson Ezra**
Department of EECS
University of Michigan
Ann Arbor, MI, USA
jezra@umich.edu

**Dragomir Radev**
Department of EECS
and School of Information
University of Michigan
Ann Arbor, MI, USA
radev@umich.edu

## Abstract

Bibliometric measures are commonly used to estimate the popularity and the impact of published research. Existing bibliometric measures provide "quantitative" indicators of how good a published paper is. This does not necessarily reflect the "quality" of the work presented in the paper. For example, when *h-index* is computed for a researcher, all incoming citations are treated equally, ignoring the fact that some of these citations might be negative. In this paper, we propose using NLP to add a "qualitative" aspect to biblometrics. We analyze the text that accompanies citations in scientific articles (which we term *citation context*). We propose supervised methods for identifying citation text and analyzing it to determine the purpose (i.e. author intention) and the polarity (i.e. author sentiment) of citation.

## 1 Introduction

An objective and fair evaluation of the impact of published research requires both quantitative and qualitative assessment. Existing bibliometric measures such as *H-Index* (Hirsch, 2005; Hirsch, 2010), *G-index* (Egghe, 2006), and *Impact Factor* (Garfield, 1994) focus on the quantitative aspect of this evaluation which dose not always correlate with the qualitative aspect.

For example, the number of papers published by a researcher only tells how productive she or he is. It does not say anything about the quality or the impact of the work. Similarly, the number of citations that a paper receives should not be used to gauge the quality of the work as it really only measures the popularity of the work and the interest of other researchers in it (Garfield, 1979). Controversial papers or those based on fabricated data or experiments may receive a large number of citations. A popular example of fraudulent research that deceived many researchers and caught media attention was the case of a South Korean research scientist, Hwang Woosuk, who was found to have faked his research results in the area of human stem cell cloning. His research was published in *Science* and received close to 200 citations after the fraud was discovered. The vast majority of those citations were negative.

This suggests that the *purpose* of citation should be taken into consideration when biblometric measures are computed. Negative citations should be weighted less than positive or neutral citations. This motivates the need to automatically distinguish between positive, negative, and neutral citations and to identify the purpose of a citation; i.e. the author's intention behind choosing a published article and citing it.

This analysis of citation purpose and polarity can be useful for many applications. For example, it can be used to build systems that help funding agencies and hiring committees at universities and research institutions evaluate researchers' work more accurately. It can also be used as a preprocessing step in systems that process scholarly data. For example, citation-based summarization systems (Qazvinian and Radev, 2008; Qazvinian et al., 2010; Abu-Jbara and Radev, 2011) and survey generation systems (Mohammad et al., 2009; Qazvinian et al., 2013) can benefit from citation purpose and polarity analysis to improve paper and content selection.

In this paper, we investigate the use of linguistic analysis techniques to automatically identify the purpose of citing a paper and the polarity of this citation. We first present a sequence labeling method for extracting the text that cites a given target reference; i.e. the text that appears in a scientific article and refers to another article and comments on it. We use the term *citation context* to refer to this text. Next,

*Proceedings of NAACL-HLT 2013*, pages 596–606,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

we use supervised classification techniques to analyze this text and identify the purpose and polarity of citation.

The rest of this paper is organized as follows. Section 2 reviews the related work. We present our approach in Section 3. We then describe the data and experiments in Section 4. Finally, Section 5 concludes the paper and suggests directions for future work.

## 2 Related Work

Our work is related to a large body of research on citations. Studying citation patterns and referencing practices has interested researchers for many years (Hodges, 1972; Garfield et al., 1984). White (2004) provides a good survey of the different research directions that study or use citations. In the following subsections, we review three lines of research that are closely related to our work.

### 2.1 Citation Context Identification

The first line of related research addresses the problem of identifying citation context. The context of a citation that cites a given target paper can be a set of sentences, one sentence, or a fragment of a sentence.

Nanba and Okumura (1999) use the term *citing area* to refer to the same concept. They define the citing area as the succession of sentences that appear around the location of a given reference in a scientific paper and have connection to it. Their algorithm starts by adding the sentence that contains the target reference as the first member sentence in the citing area. Then, they use a set of cue words and hand-crafted rules to determine whether the surrounding sentences should be added to the citing area or not. In (Nanba et al., 2000), they use their algorithm to improve citation type classification and automatic survey generation.

Qazvinian and Radev (2010) addressed a similar problem. They proposed a method based on probabilistic inference to extract non-explicit citing sentences; i.e., sentences that appear around the sentence that contains the target reference and are related to it. They showed experimentally that citation-based survey generation produces better results when using both explicit and non-explicit citing sentences rather than using the explicit ones alone.

In previous work, we addressed the issue of identifying the scope of a given target reference in citing sentences that contain multiple references (2012). Our definition of *reference scope* was limited to fragments of the explicit citing sentence (i.e. the sentence in which actual citation appears). That method does not identify related text in surrounding sentences.

In this work, we propose a supervised sequence labeling method for identifying the citation context of given reference which includes the explicit citing sentence and the related surrounding sentences.

### 2.2 Citation Purpose Classification

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

Spiegel-Rosing (1977) proposed 13 categories for citation purpose based on her analysis of the first four volumes of Science Studies. Some of them are: Cited source is the specific point of departure for the research question investigated, Cited source contains the concepts, definitions, interpretations used, Cited source contains the data used by the citing paper. Nanba and Okumura (1999) came up with a simple schema composed of only three categories: *Basis*, *Comparison*, and other *Other*. They proposed a rule-based method that uses a set of statistically selected cue words to determine the category of a citation. They used this classification as a first step for scientific paper summarization. Teufel et al. (2006), in their work on citation function classification, adopted 12 categories from Spiegel-Rosing’s taxonomy. They trained an SVM classifier and used it to label each citing sentence with exactly one category. Further, they mapped the twelve categories to four top level categories namely: weakness, contrast

(4 categories), positive (6 categories) and neutral.

The taxonomy that we use in this work is based on previous work. We adopt a scheme that contains six categories. We selected the six categories after studying all the previously used citation taxonomies. We included the ones we believed are important for improving bibliometric measures and for the applications that we are planning to pursue in the future (Section 5).

### 2.3 Citation Polarity Classification

The polarity (or sentiment) of a citation has also been studied previously. Previous work showed that positive and negative citations are common, although negative citations might be expressed indirectly or in an implicit way (Ziman, 1968; MacRoberts and MacRoberts, 1984; THOMPSON and YIYUN, 1991). Athar (2011) addressed the problem of identifying sentiment in citing sentences. He used a set of structure-based features to train a machine learning classifier using annotated data. This work uses the citing sentence only to predict sentiment. Context sentences were ignored. Athar and Teufel (2012a) observed that taking the context into consideration when judging sentiment in citations increases the number of negative citations by a factor of 3. They proposed two methods for utilizing the context. In the first method, they treat the citing sentence and a fixed context (a window of four sentences around the citing sentence) as if they were a single sentence. They extract features from the merged text and train a classifier similar to what they did in their 2011 paper. In the second method, they use a four-class annotation scheme. Each sentence in a window of four sentences around the citation is labeled as positive, negative, neutral, or excluded (unrelated to the cited work). There experiments surprisingly gave negative results and showed that classifying sentiment *without* considering the context achieves better results. They attributed this to the small size of their training data and to the noise that including the context text introduces to the data. In (Athar and Teufel, 2012b), the authors present a method for automatically identifying all the mentions of the cited paper in the citing paper. They show that considering all the mentions improves the performance of detecting sentiment in citations.

In our work, we propose a sequence labeling method for identifying the citation context first, and then use a supervised approach to determine the polarity of a given citation.

## 3 Approach

In this section, we describe our approach to three tasks: citation context identification, citation purpose classification, and citation polarity identification. We also describe a preprocessing stage that is applied to the citation text before performing any of the three tasks.

### 3.1 Preprocessing

The goal of the preprocessing stage is to clean and prepare the citation text for part-of-speech tagging and parsing. The available POS taggers and parsers are not trained on citation text. Citation text is different from normal text in that it contains references written in a special format (e.g., author names and publication year written in parentheses; or reference indices written in square brackets). Many citing sentences contain multiple references, some of which might be grouped together in a pair of parentheses and separated by a comma or a semi-colons. These references are usually not syntactic nor semantic constituents of the sentences they appear in. This results in many POS tagging and parsing errors. We address this issue in the pre-processing stage to improve the performance of the feature extraction component. We perform three pre-processing steps:

**a. Reference Tagging:** In the first step, we find and tag all the references that appear in the text. We use a regular expression to find references and replace each reference with a placeholder. The reference to the target paper is replaced by the placeholder *TREF*. Each other reference is replaced by *REF*.

**b. Reference Grouping:** In this step, we identify grouped references (i.e. multiple references listed between one pair of parentheses separated by semicolons). Each such group is replaced by a placeholder, *GREF*. If the target reference is a member of the group, we use a different placeholder: *GTREF*.

**c. Non-syntactic Reference Removal:** A reference or a group of references could either be a syntactic constituent and has a semantic role in the sentence or not (Whidby, 2012; Abu Jbara and Radev, 2012). If the reference is not a syntactic compo-

| Feature | Description |
|---|---|
| Demonstrative determiners | Takes a value of 1 if the current sentence contains contains a *demonstrative determiner* (this, these, etc.), and 0 otherwise. |
| Conjunctive adverbs | Takes a value of 1 if the current sentence starts with a *conjunctive adverb* (However, Furthermore, Accordingly, etc.), and 0 otherwise. |
| Position | Position of the current sentence with respect to the citing sentence. This feature takes one of four values: -1, 0, 1, and 2. |
| Contains Closest Noun Phrase | Takes a value of 1 if the current sentence contains closest noun phrase (if any) immediately before the reference position in the citing sentence, and 0 otherwise. This noun phrase often is the name of a method, a tool, or corpus originating from the cited reference. |
| 2-3 grams | The first bigram and trigram in the sentence (*This approach*, *One problem with*, etc.). |
| Contains Other references | Takes a value of 1 if the current sentence contains references other than the target, and 0 otherwise. |
| Contains a Mention of target reference | Takes a value of 1 if the current sentence contains a mention (explicit or anaphoric) of the target reference, and 0 otherwise. |
| Multiple references | Takes a value of 1 if the citing sentence contains multiple references, and 0 otherwise. If the citing sentence contains multiple references, it becomes less likely that the surrounding sentences are related. |

Table 1: Features used for citation context identification

nent in the sentence, we remove it to reduce parsing errors. Following our previous work (Abu Jbara and Radev, 2012), we use a rule-based algorithm to determine whether a reference should be removed from the sentence or kept. The algorithm uses stylistic and linguistic features such as the style of the reference, the position of the reference, and the surrounding words to make the decision. When a reference is removed, the head of the closest noun phrase (NP) immediately before the position of the removed reference is used as a representative of the reference. This is needed for feature extraction as shown later in the paper.

### 3.2 Citation Context Identification

The task of identifying the citation context of a given target reference can be formally defined as follows. Given a scientific article $A$ that cites another article $B$, find a set of sentences in $A$ that talk about the work done in $B$ such that at least one of these sentences contains an explicit reference to $B$.

We treat this problem as a sequence labeling problem. The goal is to find the globally best sequence of labels for all the sentences that appear within a window around the *citing sentence*. The *citing sentence* is the one that contains an explicit reference to the cited paper. Each sentence within the window is labeled as *INCLUDED* or *EXCLUDED* from the citation context of the given target paper. To determine the size of the window, we examined a development set of 300 sentences. We noticed that the related context almost always falls within a window of four sentences. The window includes the citing sentence, one sentence before the citing sentence, and two sentences after the citing sentence.

We use Conditional Random Fields (CRFs) for sequence labeling. In particular, we use a first-order chain-structured CRF. The chain consists of two sets of nodes: 1) a set of hidden nodes **Y** which represent the context labels of sentences (INCLUDED or EXCLUDED), and 2) a set of observed nodes **X** which represent the features extracted from the sentences. The task is to estimate the probability of a sequence of labels Y given the sequence of observed features X: $P(\mathbf{Y}|\mathbf{X})$

Lafferty et al. (2001) define this probability to be a normalized product of potential functions $\psi$:

$$P(\mathbf{y}|\mathbf{x}) = \prod_t \psi_k(y_t, y_{t-1}, \mathbf{x}) \tag{1}$$

Where $\psi_k(y_t, y_{t-1}, \mathbf{x})$ is defined as

$$\psi_k(y_t, y_{t-1}, \mathbf{x}) = exp(\sum_k \lambda_k f(y_t, y_{t-1}, \mathbf{x})) \tag{2}$$

where $f(y_t, y_{t-1}, \mathbf{x})$ is a transition feature function of the label at positions $i-1$ and $i$ and the observation sequence **x**; and $\lambda_j$ is a parameter that the algorithm estimates from training data.

The features we use to train the CRF model include structural and lexical features that attempt to capture indicators of relatedness to the given target reference. The features that we used and their descriptions are listed in table 1.

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

Table 2: Annotation scheme for citation purpose. Motivated by the work of (Spiegel-Rösing, 1977) and (Teufel et al., 2006)

### 3.3 Citation Purpose Classification

In this section, we describe the citation purpose classification task. Given a target paper $B$ and its citation context (extracted using the method described above) in a given article $A$, we want to determine the purpose of citing $B$ by $A$. The purpose is defined as intention behind selecting $B$ and citing it by the author of $A$ (Garfield, 1964).

We use a taxonomy that consists of six categories. We designed this taxonomy based on our study of similar taxonomies proposed in previous work. We selected the categories that we believe are more important and useful from a bibliometric point of view, and the ones that can be detected through citation text analysis. We also tried to limit the number of categories by grouping similar categories proposed in previous work under one category. The six categories, their descriptions, and an example for each category are listed in Table 2.

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

### 3.4 Citation Polarity Identification

In this section, we describe the citation polarity identification task. Given a target paper *B* and its citation context in a given article $A$, we want to determine the polarity of the citation text with respect to $B$. The polarity can be: *positive*, *negative*, or *neutral (objective)*. Positive, negative, and neutral in this context are defined in a slightly different way than their usual sense. A citation is marked positive if it either explicitly states a strength of the target paper or indicates that the work done in the target paper has been used either by the author or a third-party. It is also marked as positive if it is compared to another paper (possibly by the same authors) and deemed better in some way. A citation is marked negative if it explicitly points to a weakness of the target paper. It is also marked as negative if it is compared to another paper and deemed worse in some way. A citation is marked as neutral if it is only descriptive.

Similar to citation purpose classification, we use a supervised approach for this problem. We train a classification model using the same features listed in Table 3. Due to the high skewness in the data (more than half of the citations are neutral), we use two setups for binary classification. In the first setup, the citation is classified as *Polarized (Subjective)* or *(Neutral) Objective*. In the second one, *Subjective* citations are classified as *Positive* or *Negative*. We find that this method gives more intuitive results than using a 3-way classifier.

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

Table 3: The features used for citation purpose and polarity classification

## 4 Evaluation

In this section, we describe the data that we used for evaluation and the experiments that we conducted.

### 4.1 Data

We use the ACL Anthology Network corpus (AAN) (Radev et al., 2009; Radev et al., 2013) in our evaluation. AAN is a publicly available collection of more than 19,000 NLP papers. It includes a manually curated citation network of its papers as well as the full text of the papers and the citing sentences associated with each edge in the citation network. From this set, we selected 30 papers that have different numbers of incoming citations and that were consistently cited since they were published. These 30 papers received a total of about 3,500 citations from within AAN (average = 115 citation/paper, Min = 30, and Max = 338). These citations come from 1,493 unique papers. For each of these citations, we extracted a window of 4 sentences around the reference position. This brings the number of sentences in our dataset to a total of roughly 14,000 sentences. We refer to this dataset as *training/testing dataset*.

In addition to this dataset, we created another dataset that contains 300 citations that cite 5 papers from AAN. We refer to this dataset as the *development* dataset. This dataset was used to determine the size of the citation context window, and to develop the feature sets used in the three tasks described in Section 3 above.

### 4.2 Annotation

In this section, we describe the annotation process. We asked graduate students with good background in NLP (the topic of the annotated sentences) to provide three annotations for each citation example (a window of 4 sentences around the reference anchor) in the *training/testing dataset*. We asked them to mark the sentences that are related to a given target reference. In addition, we asked them to determine the purpose of citing the target reference by choosing from the six purpose categories that we described earlier. We also asked them to determine whether the citation is negative, positive, or neutral.

To estimate the inter-annotator agreement, we picked 400 sentences from the *training/testing dataset* and assigned them to two different annotators. We use the Kappa coefficient (Cohen, 1968) to measure the agreement. The Kappa coefficient is defined as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \qquad (3)$$

where P(A) is the relative observed agreement among annotators and P(E) is the hypothetical probability of chance agreement. The agreement between the two annotators on the context identification task was $K = 0.89$. On Landis and Kochs (Landis and Koch, 1977) scale, this value indicates *almost perfect* agreement. The agreement on the purpose and the polarity classification task were $K = 0.61$ and $K = 0.66$, respectively; which indicates *substantial agreement* on the same scale.

The annotation shows that in 22% of the citation examples, the citation context consists of 2 or more sentences. The distribution of the purpose categories in the data was: 14.7% criticism, 8.5% comparison, 17.7% use, 7% substantiation, 5% basis, and 47% other. The distribution of the polarity categories was: 30% positive, 12% negative, and 58% neutral.

### 4.3 Experimental Setup

We use the CRF++[1] toolkit for CRF training and testing. We use the Stanford parser to parse the citation text and generate the dependency parse trees of sentences. We use Weka for classification experiments. We experimented with several classifiers including: SVM, Logistic Regression (LR), and Naive Bayes. All the experiments that we conducted used the *training/testing dataset* in a 10-fold cross validation mode. All the results have been tested for statistical significance using a 2-tailed paired t-test.

### 4.4 Evaluation of Citation Context Identification

We compare the CRF approach to three baselines. The first baseline (ALL) labels all the sentences in the citation window of size 4 as *INCLUDED* in the citation context. The second baseline (CS-ONLY) labels the citing sentence only as *INCLUDED* in the citation context. In the third baseline, we use a supervised classification method instead of sequence labeling. We use Support Vector Machines (SVM) to train a model using the same set of features as in the CRF approach.

Table 4 shows the precision, recall, and F1 score of the CRF approach and the baselines. The results show that our CRF approach outperforms all the baselines. It also asserts our expectation that addressing this problem as a sequence labeling problem leads to better performance than individual sentence classification, which is also clear from the nature of the task.

| | Precision | Recall | F1 |
|---|---|---|---|
| CRFs | **98.5%** | **82.0%** | **89.5%** |
| ALL | 30.7% | 100.0% | 46.9% |
| CS-ONLY | 88.0% | 74.0% | 80.4% |
| SVM | 92.0% | 76.4% | 83.5% |

Table 4: Results of citation context identification

**Feature Analysis:** We evaluated the importance of the features listed in Table 1 by computing the chi-squared statistic for every feature with respect to the class. We found that the lexical features (such as determiners and conjunction adverbs) are generally more important than the structural features (such as position and reference count). The features shown in Table 1 are listed in the order of their importance based on this analysis.

### 4.5 Evaluation of Citation Purpose Classification

Our experiments with several classification algorithms showed that the SVM classifier outperforms Logistic Regression and Naive Bayes classifiers. Due to space limitations, we only show the results for SVM. Table 5 shows the precision, recall, and F1 for each of the six categories. It also shows the overall accuracy and the Macro-F measure.

**Feature Analysis:** The chi-squared evaluation of the features listed in Table 3 shows that both lexical and structural features are important. It also shows that among lexical features, the ones that are limited to the existence of a direct relation to the target reference (such as *closest* verb, adjective, adverb, subjective cue, etc.) are most useful. This can be explained by the fact that the restricting the features to having direct dependency relation introduces much less noise than other features (such as *Dependency Triplets*). Among the structural features, the number of references in the citation context showed to be more useful.

### 4.6 Evaluation of Citation Polarity Identification

Similar to the case of citation purpose classification, our experiments showed that the SVM classifier outperforms the other classifiers that we experimented with. Table 6 shows the precision, recall, and F1 for

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

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

Table 5: Summary of Citation Purpose Classification Results (10-fold cross validation, SVM: Linear Kernel, c = 1.0)

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

### 4.7 Impact of Context on Classification Accuracy

To study the impact of using citation context in addition to the citing sentence on classification performance, we ran two polarity classification experiments. In the first experiment, we used the citing sentence only to extract the features that are used to train the classifiers. In the second experiment, we used the gold context sentences (the ones labeled *INCLUDED* by human annotators). Table 6 shows the results of the first experiment between rounded parentheses and the results of the second experiments in square brackets. The results show that adding citation context improves the classification accuracy especially in the *subjective* categories, specially in the negative category if we want to be more specific. This supports our intuition about polarized citations that authors start their review of the cited work with an objective (neutral) sentence and then follow it with their criticism if they have any. We also reached to similar conclusions with purpose classification, but we are not showing the numbers due to space limitations.

### 4.8 Other Experiments

#### 4.8.1 Can We Do Better?

In this section, we investigate whether it is possible to improve the performance in the two classification tasks. One factor that we believe could have an impact on the result is the size of the training data. To examine this hypothesis, we ran the experiment on different sizes of data. Figure 1 shows the learning curve of the two classifiers for different sizes of training data. The accuracy increases as more training data is available so we can expect that with even more data, we can do even better.

| | Negative % | Positive % | Neutral % |
|---|---|---|---|
| Precision | 68.7 (66.4) [69.8] | 54.9 (52.1) [55.4] | 83.6 (82.8) [84.2] |
| Recall | 79.2 (71.1) [81.1] | 48.1 (45.6) [46.3] | 95.5 (95.1) [95.3] |
| F1 | 73.6 (68.7) [75.0] | 51.3 (48.6) [50.4] | 89.1 (88.5) [89.4] |
| Accuracy: 81.4 (74.2) [84.2] % | | | |
| Macro-F: 71.3 (62.1) [74.2] % | | | |

Table 6: Summary of Citation Polarity Classification Results (10-fold cross validation, SVM: Linear Kernel, c = 1.0). Numbers between rounded parentheses are when only the explicit citing sentence is used (i.e. no context). Numbers in square brackets are when the gold standard context is used.

#### 4.8.2 Relation Between Citation Purpose/Polarity and Citation Count

The main motivation of this work is our hypothetical assumption that using NLP for analyzing citations gives a clearer picture of the impact of the cited work. As a way to check the validity of this assumption, we study the correlation between the counts of the different purpose and polarity categories. We also study the correlation between these categories and the total number of citations that a paper received since it was published. We use the *training/testing dataset* and the gold annotations for this study.

We compute the Pearson correlation coefficient between the counts of citations from the different categories that a paper received per year since its publication. We found that, on average, the correlation between positive and negative citations is negative (AVG P = -0.194) and that the correlation be-

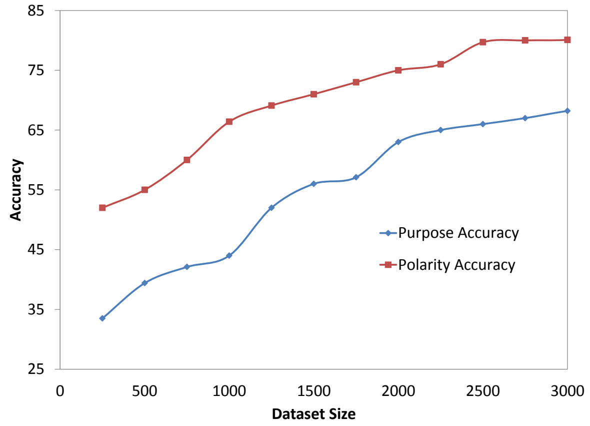

Figure 1: The effect of size of the data set size on the classifiers accuracy.

tween the count of positive citations and the total number of citations is higher than the correlation between negative citations and total citations (AVG P = 0.531 for positive vs. AVG P = 0.054 for negative).

Similarly, we noticed that there is a higher positive correlation between *Use* citations and total citations than in the case of both *Substantiation* and *Basis*. This can be explained by the intuition that publications that present new algorithms, tools, or corpora that are used by the research community become more and more popular with time and thus receive more and more citations.

Figure 2 shows the result of running our purpose classifier on all the citations to Papineni et al.'s (2002) paper about Bleu, an automatic metric for evaluating Machine Translation (MT) systems. The figure shows that this paper receives a high number of *Use* citations. This makes sense for a paper that describes an evaluation metric that has been widely used in the MT area. The figure also shows that in the recent years, this metric started to receive some *Criticizing* citations that resulted in a slight decrease in the number of *Use* citations. Such a temporal analysis of citation purpose and polarity is useful for studying the dynamics of research. It can also be used to detect the emergence or de-emergence of research techniques.

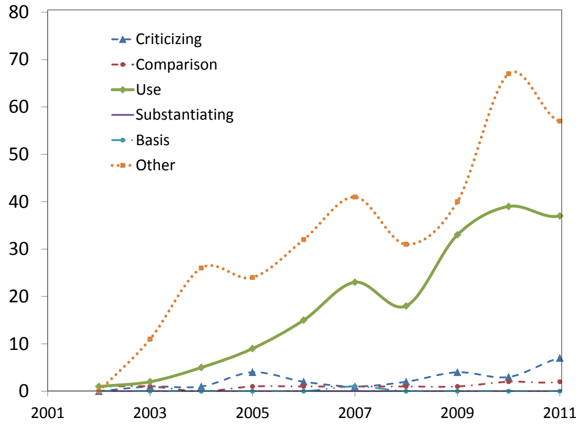

Figure 2: Change in the purpose of the citations to Papineni et al. (2002)

## 5 Conclusion

In this paper, we presented methods for three tasks: citation context identification, citation purpose classification, and citation polarity classification. This work is motivated by the need for more accurate bibliometric measures that evaluates the impact of research both qualitatively and quantitatively. Our experiments showed that we can classify the purpose and polarity of citation with a good accuracy. It also showed that using the citation context improves the classification accuracy and increases the number of polarized citations detected. For future work, we plan to use the output of this research in several applications such as predicting future prominence of publications, studying the dynamics of research, and designing more accurate bibliometric measures.

## Acknowledgement

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

## References

Daryl E. and Soumyo D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5(4):pp. 423–441.

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 500–509, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

Awais Athar and Simone Teufel. 2012a. 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, Montréal, Canada, June. Association for Computational Linguistics.

Awais Athar and Simone Teufel. 2012b. Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 18–26, Jeju Island, Korea, July. Association for Computational Linguistics.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA, June. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Susan Bonzi. 1982. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, 33(4):208–216.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Leo Egghe. 2006. Theory and practise of the g-index. *Scientometrics*, 69:131–152.

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

Eugene Garfield. 1964. Can citation indexing be automated?

E. Garfield. 1979. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4):359–375.

Eugene Garfield. 1994. The thomson reuters impact factor.

J. E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, November.

J. E. Hirsch. 2010. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, December.

T. L. Hodges. 1972. Citation indexing-its theory and application in science, technology, and humanities. *Ph.D. thesis, University of California at Berkeley.Ph.D. thesis, University of California at Berkeley.*

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, March.

Michael H. MacRoberts and Barbara R. MacRoberts. 1984. The negational reference: Or the art of dissembling. *Social Studies of Science*, 14(1):pp. 91–94.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 584–592, Boulder, Colorado, June. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2012. *sem 2012 shared task: resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 265–274, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *IJCAI '99: Proceedings of the Six-*

*teenth International Joint Conference on Artificial Intelligence*, pages 926–931, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hidetsugu Nanba, Noriko Kando, Manabu Okumura, and Of Information Science. 2000. Classification of research papers using citation links and citation types: Towards automatic review article generation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696, Manchester, UK, August. Coling 2008 Organizing Committee.

Vahed Qazvinian and Dragomir R. Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 555–564, Uppsala, Sweden, July. Association for Computational Linguistics.

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Ozgur. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 895–903, Beijing, China, August. Coling 2010 Organizing Committee.

Vahed Qazvinian, Dragomir R. Radev, Saif Mohammad, Bonnie Dorr, David Zajic, Michael Whidby, and Taesun Moon. 2013. Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research*.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The acl anthology network corpus. In *NLPIR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 54–61, Morristown, NJ, USA. Association for Computational Linguistics.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, pages 1–26.

Ina Spiegel-Rösing. 1977. Science Studies: Bibliometric and Content Analysis. *Social Studies of Science*, 7(1):97–113, February.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *In Proc. of EMNLP-06*.

GEOFF THOMPSON and YE YIYUN. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, 12(4):365–382.

Melvin Weinstock. 1971. *Citation Indexes*. Encyclopedia of Library and Information Science.

Michael Alan Whidby. 2012. Citation handling: Processing citation text in scientific documents. In *Master Thesis*.

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

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. M. Ziman. 1968. *Public knowledge: An essay concerning the social dimension of science*. Cambridge U.P., London.