Towards Fine-grained Citation Function Classification

Xiang Li **Yifan He** **Adam Meyers** **Ralph Grishman**
Computer Science Department
New York University
{xiangli, yhe, meyers, grishman}@cs.nyu.edu

Abstract

We look into the problem of recognizing citation functions in scientific literature, trying to reveal authors' rationale for citing a particular article. We introduce an annotation scheme to annotate citation functions in scientific papers with coarse-to-fine-grained categories, where the coarse-grained annotation roughly corresponds to citation sentiment and the fine-grained annotation reveals more about citation functions. We implement a Maximum Entropy-based system trained on annotated data under this scheme to automatically classify citation functions in scientific literature. Using combined lexical and syntactic features, our system achieves the F-measure of 67%.

1 Introduction

Citations in scientific papers serve different purposes, from comparing one work to another to acknowledging the inventor of certain concepts. Recognizing citation functions is important for understanding the structure of a single scientific document as well as mining citation graphs within a document collection. Therefore, this task has attracted researchers from the fields of discourse analysis, sociology of science, and information sciences for decades (Teufel et al., 2006a).

Most of the existing research in this area focused on the analysis of citation sentiment, which has achieved good accuracy (see, e.g., (Teufel et al., 2006a)). Citation sentiment analysis systems are usually able to identify positive, neutral, or negative opinions, but if we want to better understand the exact function of a citation, we need to know not only whether the authors like the citation, but also how the citation is used in a given context (Section 2).

In this paper, we try to reveal citation functions more accurately than simply classifying citation sentiment. We first create a two level coarse-to-fine grained annotation scheme (Section 3). The coarse-level annotation corresponds roughly to sentiment categories, including `POSITIVE`, `NEGATIVE`, and `NEUTRAL`. The fine-grained annotation scheme provides a more detailed description of citation functions, such as `Significant`, which asserts the importance of an article or a work, and `Discover`, which acknowledges the original discoverer/inventor of a method or material.

Using data annotated under this scheme, we train classifiers to determine citation functions, and experiment with features from lexical to syntactic levels (Section 4). We predict the fine-grained citation function at 67% in F-measure in our experiments, which is at the same level as the coarse-grained citation sentiment classification (Section 5).

2 Related Work

The background for our work is in citation analysis. Applications of citation analysis include evaluating the impact of a published literature through a measurable bibliometric (Garfield, 1972; Luukkonen, 1992; Borgman and Furner, 2002), analyzing bibliometric networks (Radev et al., 2009), summarizing scientific papers (Qazvinian and Radev, 2008; Abu-Jbara and Radev, 2011), generating surveys of scientific paradigms (Mohammad et al., 2009), among others. Correctly and accurately recognizing citation functions is a cornerstone for these tasks.

Proceedings of Recent Advances in Natural Language Processing, pages 402–407,
Hissar, Bulgaria, 7-13 September 2013.

Citation Function	Description
Based_on^{+}	A work is based on the cited work
Corroboration^{+}	Two works corroborate each other
Discover^{+}	Acknowledge the invention of a technique
Positive^{+}	The cited work is successful
Practical^{+}	The cited work has a practical use
Significant^{+}	The cited work is important
Standard^{+}	The cited work is a standard
Supply^{+}	Acknowledge the supplier of a material
Contrast$^{=}$	Compares two works in a neutral way
Co-citation$^{=}$	Citations that appear closely
Neutral$^{=}$	The cited work not belonging to other functions
Negative^{-}	The weakness of the cited work is discussed

Table 1: Annotation Scheme for Citation Function: $^{+}$ represents POSITIVE sentiment, $^{=}$ represents NEUTRAL sentiment, and $^{-}$ represents negative sentiment

Researchers have introduced several annotation schemes for citation analysis. The work of Teufel et al. (2006b) is the most related to ours. They proposed an annotation scheme for citation functions based on why authors cite a particular paper, following Spiegel-Rüsing (1977). This scheme provides clear definition for some of the basic citation functions, such as Contrast, but mainly concerns the citations that authors compare to or build upon, ignoring the relationship between two cited works. Sometimes the relationship between two cited works is also meaningful and important, from which we can know more about the functions and influences of one cited work on other works. For example, the cited work may be utilized or applied by another cited work, which would be captured by Practical in our annotation scheme but considered as neutral under their scheme. In addition, their annotation scheme does not explicitly recognize milestone or standard work in a particular research field, while our annotation scheme does through the Significant function. We continue to use these basic functions, but try to expand their scheme by incorporating more functions, such as acknowledgement and corroboration, which reflects the attitude of the research community towards a citation.

Regarding the automatic recognition of citation functions or citation categories, Teufel et al. (2006a) presented a supervised learning framework to classify citation functions mainly utilizing features from cue phrases. Athar (2011) explored the effectiveness of sentence structure-based features to identify sentiment polarity of citations. Dong and Schäfer (2011) proposed a four-category definition of citation functions following Moravcsik and Murugesan (1975) and a self-training-based classification model. Different from previous work that mainly classified citations into sentiment categories or coarse-grained functions, our scheme, we believe, is more fine-grained. It is also worth noting that Teufel et al. (2006a), Athar (2011), and Dong and Schäfer (2011) all worked on citations in computational linguistics papers, but we investigate citations in biomedical articles.

3 Annotation

Our annotation scheme contains three general citation function categories POSITIVE, NEUTRAL, and NEGATIVE: POSITIVE citations reflect agreement, usage, or compatibility with cited work; NEUTRAL citations refer to related knowledge or background in cited work; and NEGATIVE citations show weakness of cited work. These three general categories are often used as citation sentiments in previous citation sentiment analysis work. We extend these categories by sorting them into smaller subcategories that reflect the functions of citations. POSITIVE (see $^{+}$ in Table 1), for example, shows a general sentiment of agreement. We divide POSITIVE into Based_on, Corroboration, Discover, Positive, Practical, Significant, Standard, and Supply in order to more accurately describe how a citation is used. The details about each citation function are summarized in Table 1. We provide

Citation Function	Example
Based_on^{+}	*Results* ***based on*** *the Comparative Toxicogenomics Database (CTD)* ***[14]****, we constructed a human P-PAN.*
Corroboration^{+}	*This observation is* ***in accordance with*** *previously published data* ***[39]****.*
Discover^{+}	*The core of our procedure is derived from the "target hopping" concept* ***defined*** *previously* ***[3]****.*
Positive^{+}	*Therefore, a systems biology approach, such as the one that* ***was successfully employed*** *by Chen and colleagues* ***[1]****, is an effective alternative for analyzing complex diseases.*
Practical^{+}	*Molecular Modeling and Docking Genetic algorithm GOLD (Genetic Optimization for Ligand Docking), a docking program based on genetic algorithm* ***[39][42] was used to*** *dock the ligands to the protein active sites.*
Significant^{+}	*In addition to nanomaterial composition, size and concentration, the influence of cell type* ***is of paramount importance*** *in nanomaterial toxicity as highlighted in other recent investigations in cell vs. cell comparisons* ***[49]****.*
Standard^{+}	*A* ***standard*** *genetic algorithm* ***[31]*** *was used to select the final physicochemical properties of Pafig with population size of 10, crossover probability of 0.8, mutation probability of 0.01 and predetermined number of 200 generations.*
Supply^{+}	*The rate constants* ***obtained directly from*** *the ultrafast, time-resolved optical spectroscopic experiments carried out* ***(Polivka et al. 2005)*** *are shown in Table.*
Contrast$^{=}$	***In contrast to Rodgers et al., [34]*** *who targeted planktonic species in AMD solutions and sediments,* ***Bond et al. [37]*** *primarily sampled biofilms.*
Co-Citation$^{=}$	*They bear specific regulatory properties and mechanisms* ***(Babu et al, 2004; Wang and Purisima, 2005)****.*
Neutral$^{=}$	***Lage and collaborators [12]*** *predicted 113 new disease-candidate genes by comparing their protein-interaction* neighborhood with associated phenotypes.
Negative^{-}	*A range of methods have been applied to S. mutans typing, one of the earliest of which was based on susceptibilite to bacteriocins* ***[14], [15]*** *but was found to* ***lack reproducibility and was not readily transferred between laboratories****.*

Table 2: Citation Function Examples

an example for each function in Table 2 to illustrate how it is defined.

Two annotators are trained to perform the annotation. The articles we work on are from the open access subset of PubMed, which consists of articles from the biomedical domain. We require the annotators to mark citation functions, and point to textual evidence for assigning a particular function.

4 Recognizing Citation Functions

We use the Maximum Entropy (MaxEnt) model to classify all citations into the above citation function categories. We experiment with both surface and syntactic features. When parsing the context sentence, we replace each citation content with a <CITATION> symbol, in order to remove the contextual bias.

4.1 Surface Features

We capture n-grams, signal words collected by system developers, pronouns, negation words, and words related to formulae, graphs, or tables in the context sentence as surface level features.

- **N-Gram Features** use both uni-grams of the context sentence and the tri-gram context window that contains the citation.
- **Signal Word Features** check whether the text signals for a citation function (151 words/phrases in total, collected by system developers from dictionaries) appear in the context sentence.
- **Pronoun Features** look for third-person pronouns and their positions in the context sentence.

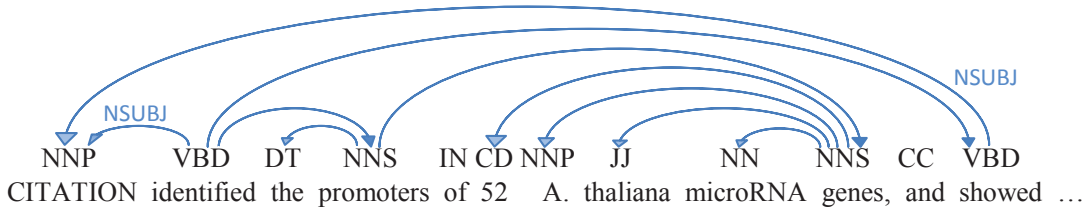

Figure 1: POS and Dependency Features

- **Negation Features** fire if negation words (135 words in total) appear in the context sentence with its scope.
- **FGT Features** fire if words or structures like formula, graph, or table appear in the context sentence.

4.2 Syntactic Features

We capture more generalized or long-distance information by taking advantages of syntactic features.

The Part-of-Speech Features use Part-of-Speech (POS) tags adds generalizability to surface level signals, e.g., "VERB with" covers signals like "experiment with" and "solve with", which might indicate a `Practical` function. We use a combination of POS tags and words in a two-word context window around the <CITATION> as features. In Figure 1, "VBD_DT", "identified_DT", and "VBD_the" would be extracted.

The Dependency Features use the dependency structure of the context sentence to capture grammatical relationships between a citation and its signal words regardless of the distance between them. We extract both dependency triples and dependency labels as features. In Figure 1, if we extract dependency relations and labels attached to a <CITATION>, we would obtain "NSUBJ_identified_CITATION", "NSUBJ", and "NSUBJ_showed_CITATION" as dependency features. "NSUBJ_showed_CITATION" captures the long-distance relation between <CITATION> and a signal word "showed", which other features miss.

5 Experiments

From 91 annotated articles with total 6,355 citation instances, we train our model and test the performance through a 10-fold cross-validation procedure, so that each fold randomly contains 9 (or 10) articles with their associated citation instances.

Features	P	R	F1
baseline	0.67	0.44	0.53
baseline + fgt	0.67	0.44	0.53
baseline + sig	0.67	0.44	0.53
baseline + neg	0.68	0.44	0.54
baseline + pron	0.68	0.44	0.54
baseline + dep	0.72	0.54	0.62
baseline + pos	**0.75**	0.58	0.65
baseline + pos + dep	0.74	**0.61**	**0.67**

Table 3: Overall Performance Using Different Features: n-gram features (baseline), FGT features (fgt), signal word features (sig), negation features (neg), pronoun features (pron), dependency structure features (dep), and Part-of-Speech features (pos).

Table 3 shows the overall performance in Precision (P), Recall (R), and F-measure (F1) by incorporating different feature sets, at a 99.8% confidence level according to the Wilcoxon Matched-Pairs Signed-Ranks Significance Test. If we randomly assign one of the citation function classes to each citation instance, the performance is only 3.8% in F-measure. In addition, a simple majority classifier assigns each citation with whichever class that is in the majority in the training set, also only obtaining F-measure of 42.2%. Our results clearly show that our MaxEnt system easily outperforms these two simple baseline classifiers.

We report macro-average numbers over all citation functions, except for `NEUTRAL:Neutral`, which simply reflects that a work is cited without any particular information. We observe that surface features do not work well enough alone, as they cannot generalize beyond the signal knowledge observed in a relatively small training set. Syntactic features, on the other hand, can utilize linguistic knowledge to solve the problem, and lead to better results.

We compare F-measure of coarse-grained sentiment classification and fine-grained citation func-

Function Class	P	R	F1	Distribution
Based_on^{+}	0.250	0.029	0.051	0.028
Corroboration^{+}	**1.000**	0.022	0.043	0.036
Discover^{+}	0.861	0.750	**0.802**	0.123
Positive^{+}	N/A	0.000	N/A	0.001
Practical^{+}	N/A	0.000	N/A	0.010
Significant^{+}	N/A	0.000	N/A	0.006
Standard^{+}	0.500	0.333	0.400	0.002
Supply^{+}	0.000	0.000	N/A	0.012
Contrast$^{=}$	0.667	0.250	0.364	0.006
Co-Citation$^{=}$	0.721	**0.792**	0.755	0.333

Table 4: Performance and Distribution of Citation Function Classes

Citation Sentiment	P	R	F1
coarse-grained POSITIVE	0.93	0.45	0.60
fine-grained POSITIVE	0.82	0.43	0.57

Table 5: Comparison of Coarse- and Fine-grained Citation Function Classification on POSITIVE

tion prediction on more interesting POSITIVE functions in Table 5. We see that coarse-grained classification performs only slightly better. We suspect that each citation function in the POSITIVE category needs different signal information to identify, so a more fine-grained annotation scheme could lead to a stronger correlation between a class label and its signals. This can explain the close performance between these two paradigms, although citation function prediction is more informative and harder.

We report performance and distribution in annotated data for each citation function in Table 4. Note that the numbers in the "Distribution" column does not sum to 1, because we omit the NEUTRAL:Neutral category that does not carry information and some categories (e.g., Negative) that are too few (e.g., less than 5) in the corpus. We see that some of the functions (such as Discover) can perform much better than others. The major reason for the difference in performance is the imbalance distribution of citation functions in the annotated corpus, which, in turn, results in the difference in prediction ability of our classifier. In the extreme case, our system fails to find any positive instance for some of the categories because of the scarcity of training examples. In order to mitigate this problem, we plan to perform more function-specific annotation to obtain more data on current scarce functions.

6 Conclusion

In this paper, we introduced the task of citation sentiment analysis and citation function classification, which aims to analyze the fine-grained utility of citations in scientific documents. We described an annotation scheme to annotate citation functions in scientific papers into fine-grained categories. We presented our Maximum Entropy-based system to automatically classify the citation functions, explored the advantages of different feature sets, and confirmed the necessity of using syntactic features in our task, obtaining 67% of final F-measure score.

For future work, we plan to explore more features and perform more citation function-specific annotation for scarce functions in the current annotated corpus. Furthermore, we will also apply our annotation scheme and classification method in scientific literature from different domains, as well as investigate more elaborate machine learning models and techniques.

Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. 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. Approved for Public Release; Distribution Unlimited.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. *Coherent Citation-Based Summarization of Scientific Papers*. In Proceedings of ACL 2011, Portland, Oregon, USA.

Awais Athar. 2012. *Sentiment Analysis of Citations using Sentence Structure-Based Features*. In Proceedings of ACL-HLT 2011 Student Session, Portland, Oregon, USA.

Awais Athar and Simone Teufel. 2012. *Detection of Implicit Citations for Sentiment Detection*. In Proceedings of ACL 2012 Workshop on Discovering Structure in Scholarly Discourse, Jeju Island, South Korea.

Awais Athar and Simone Teufel. 2012. *Context-Enhanced Citation Sentiment Detection*. In Proceedings of NAACL-HLT 2012, Montreal, Canada.

Christine Borgman and Jonathan Furner. 2002. *Scholarly Communication and Bibliometrics*. Annual Review of Information Science and Technology: Vol. 36.

Jean Carletta. 1996. *Assessing Agreement on Classification Tasks: The Kappa Statistic*. Computational Linguistics, 22(2):249254.

Cailing Dong and Ulrich Schäfer. 2011. *Ensemble-style Self-training on Citation Classification*. In Proceedings of IJCNLP 2011, Chiang Mai, Thailand.

Eugene Garfield. 1955. *Citation Indexes for Science - A New Dimension in Documentation through Association of Ideas*. Science, July 15, 1955: 108-111.

Eugene Garfield. 1965. *Can Citation Indexing Be Automated?*. Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington.

Eugene Garfield. 1972. *Citation Analysis as a Tool in Journal Evaluation*. Essays of an Information Scienties, Vol 1, p.527-544.

Terttu Luukkonen. 1992. *Is Scientists' Publishing Behaviour Reward-seeking?*. Scientometrics, 24: 297319.

Saif Mohammad, Boonie 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 NAACL-HLT 2009, Boulder, Colorado, USA.

Michael J. Moravcsik and Poovanalingam Murugesan. 1975. *Some Results on the Function and Quality of Citations*. Social Studies of Science, 5:8692.

Vahed Qazvinian and Dragomir R. Radev. 2008. *Scientific Paper Summarization Using Citation Summary Networks*. In Proceedings of Coling 2008, Manchester, UK.

Dragomir R. Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2009. *A Bibliometric and Network Analysis of the field of Computational Linguistics*. Journal of the American Society for Information Science and Technology.

Ina Spiegel-Rüsing. 1977. *Bibliometric and Content Analysis*. Social Studies of Science, 7:97113.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. *Automatic classification of citation function*. In Proceedings of EMNLP 2006, Sydney, Australia.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. *An annotation scheme for citation function*. In Proceedings of Sigdial 2006, Sydney, Australia.