Automatic Citation Function Analysis with Rich Linguistic Features

Abstract
Researchers cite for different purposes – some for laying study background while some others for comparisons. This drives us to study the function roles for citations in academic publications. Although existing research has made many attempts to develop automated algorithm for large-scale analysis, continuous performance improvement is still helpful, which is the focus of this poster. We observe that the linguistic features for the citing content (the surrounding content when citing) are important to determine such role while it is often overlooked in other studies. Therefore, we are interested in understanding how different linguistic features (e.g., functional words, syntactic features) can help on improving algorithm performance. Our experiments, based on an existing dataset, shows that these features can contribute to improve existing study by 20%.

Keywords: Citation Function; Citation Function Classification; Citation Analysis; Citation Content; Natural Language Processing
Citation: Editor will add citation

Acknowledgements: If there is an acknowledgement add it here. This field is optional. For review purpose, do not add information to this field until after the paper is accepted.
Research Data: If there is any research data / raw data contact the editor. The editor will add research data information such as a DOI. This field is optional.
Contact: Author will add e-mail address.

1 Introduction

Traditional citation analysis is often criticized for oversimplifying the authors' real citing motivation by assuming a linear and equal relationship for all citations. The citation content contains abundant semantic information, which were used to enrich the presentation of classic citation network-based analysis as well as improve citation-based applications like academic influence evaluation (Moed, 2006), summarization (Qazvinian & Radev, 2008) and literature retrieval(Liu et al., 2013). Many researchers are aware of the importance of different citation functions and have applied this idea for analyzing citation sentiment(Athar, 2011), identifying citation function(Teufel, Siddharthan, & Tidhar, 2006) and citation importance(Wan & Liu, 2014) et al. Among them, citation function is considered to be the most important component as it presents different role of citation in scientific literature, from introducing related research background to acknowledging the important ideas used in current paper, which is meaningful for improving citation analysis and academic applications.

Due to the high labor cost for manual annotation, researchers started to developing automatic citation function classification algorithms. For example, Garzone (1997) built a rule-based classifier based on his self-proposed scheme, consisting of 35 different categories. Teufel et al. (2006) trained a classifier using the IBk algorithm based on a modified classification scheme containing 12 categories. Radoulov (2008) reduced the number of citation function categories by describing them as a combination of citing reason and object. Instead of extracting features automatically, he consulted linguistic expert to find useful lexical and syntactic features. Dong and Schäfer (2011) utilized the textual, physical and syntactic features, and a semi-supervised algorithm was introduced to extend the small training dataset in citation classification by making use of unlabeled data. Jochim and Schütze (2012) provided a comprehensive exploration of features used in previous studies and introduced several novel features.

One important disadvantage of the existing studies is that the classification performance still has much room to improve. Dong et al.(Dong & Schäfer, 2011) achieved 0.67 of macro F-score on 4 categories, other studies with more classification categories generally performed even worse (Abu-Jbara & Radev, 2012; Teufel et al., 2006). To further improve the classification performance and achieve reasonable performance, we proposed some new lexical and syntactic features. Our experimental results show the effectiveness of these novel features, with 86.54% accuracy and a macro F-score of 0.79, which outperforms the current state-of-art algorithm by 20%.

2 Methodology

2.1 Dataset and Classification Scheme

Most citation function analysis studies conducted automatic classification based on self-proposed classification scheme and self-annotated corpus. In this study, however, we adopt the datasets and classification scheme in Dong and Schäfer's (2001) for fair result comparison. Their scheme covers most general citation functions and mutually exclusive categories. The four categories are as follows:

- Background: research background of current work.
- Fundamental idea (Idea for short): previous work inspired the current work.
- Technical basis (Basis for short): important tools, methods, data and other resources used or adapted in the current work.
- Comparison: citing for comparing the methods or results.

The dataset contains 1,768 annotated citations that are extracted from 122 conference papers of ACL Anthology. And the number of citations in each category is 1,150 (background), 421 (Idea), 127(Basis) and 70 (Comparison).

2.2 Features

The features we used in this study consist of both the features used in existing studies and our proposed novel features. They are grouped into four categories.

- Word-level features. We use three types of word features from previous research: (1) n-gram word feature introduced by (Athar, 2011), where we only consider the unigram feature since it already captured the key lexical information without introducing too much noise; (2) cue words used by Dong et al.(Dong & Schäfer, 2011), especially the subject cue that can distinguish the informative categories from Background citations significantly; (3) modality words, main verb and root verb introduced by (Jochim & Schütze, 2012).

  Here, we added two features by observing that digits and percentages occur frequently when citing for Comparison, and words denote future work commonly occur in Background category. Thus two Boolean values are included in our feature set to indicate whether the citing sentence contains the digits and words of "Future work".
- Syntactic features. Four kinds of previous syntactic features are used here: (1) dependency relations which showed notable improvement by (Athar, 2011) for citation sentiment classification; (2) seven types of syntactic patterns captured by Dong and Schäfer (2011); (3) several detail features designed by Jochim and Schütze (2012), such as whether the citation is labeled as a constituent in sentence, whether a pronoun is linked to a comparative; (4) signal words linking to citation marks, introduced by Abu-Jbara and Radev (2012) and Li, He, Meyers, and Grishman (2013).

  Inspired by the subject cue feature from (Dong & Schäfer, 2011), we find that the verbs and adjectives linked to subject cue (the first pronoun refers to the author) play significant roles in recognizing citation function. To illustrate this idea better, figure 1 shows the dependency relations in a citing sentence. The function of citation ("Hillard et al.") here is Comparison and this function can be easily recognized with the verb and adjective cues ("obtain" and "better"). We found that these cues usually are subject to the first personal pronoun ("We" here). Thus we extract these words linked to first personal pronoun out as important features.

Figure 1. Example shows the key verb and adjective connected to first pronoun

- Physical features. This feature set contains the position and frequency information of each citation. (1) section position, mapping the citation located section into six predefined categories (Introduction, Related work, Method, Experiment, Evaluation and Conclusion) by (Dong & Schäfer, 2011) ; (2) the number of other citations in the citation sentence, an effective feature to see the importance of this citation.
- Other features: (1) self-citing feature, firstly introduced by (Teufel et al., 2006) which assumes that self-citing may indicate important citing relations; (2) named-entity recognition, used to find

whether the citation is related to resource or tool. We extracted this feature by cue words instead of building a NER (Name entity recognition) tagger.

3 Experimental Results

In this section we compare our performance with five baselines. The first baseline uses only the unigram features (see (1) in Word-level features) and the second baseline employs both the unigram word feature and the dependency relations (see (Athar, 2011)). The purpose of the above two baseline experiments is to see what the performance would be on simplest text features. We also re-implement the feature sets proposed by (Dong & Schäfer, 2011), (Abu-Jbara & Radev, 2012) and (Jochim & Schütze, 2012) as another three baselines for comparison. Our feature set includes all the features discussed in section 2.2. All the experiments are conducted through Support Vector Machine classifier with RBF kernel on optimal parameters and in a 10-fold cross validation. Besides, a feature selection based on information gain is conducted in order to select the best features.

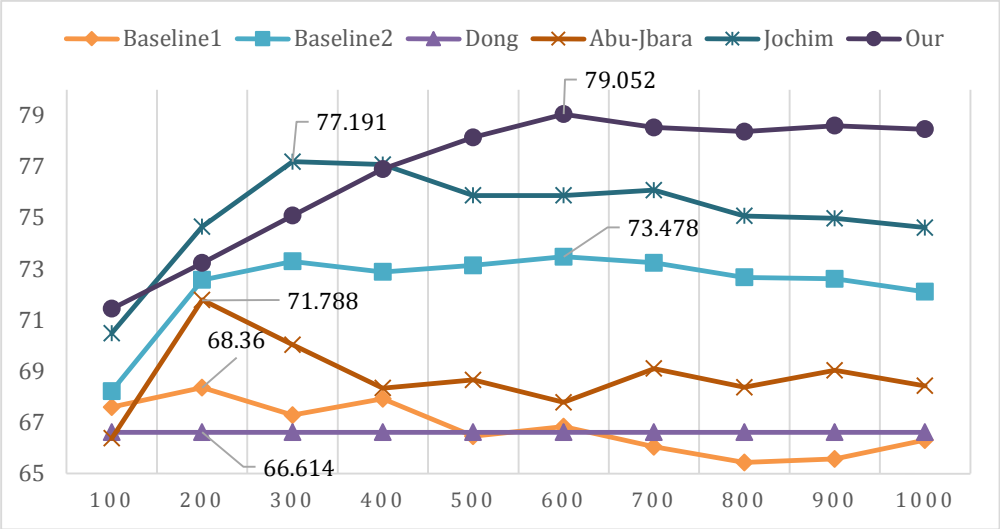

Figure 2. Macro F-score of each feature set on different size of feature selection (Y: Macro F-score, the bigger the better; X: number of features selected by Information Gain)

Macro F-score that leverages both precision and recall is preferred for measuring model performance instead of accuracy as there are skewed number of examples in each class. Figure 2 shows the classification performance for different approaches, where we have several interesting findings: first, we can see that only a small number of features (several hundred out of total 30,000 features) are useful for our task. Second, the first two baseline experiments already performed reasonable well. This tells us sometimes simple textual features are enough to yield good performance on citation classification. Third, we find that the feature sets of (Dong & Schäfer, 2011) and (Abu-Jbara & Radev, 2012) failed as they mainly based on hand-crafted cue words, which unable to handle the high complexity in real situation and in different data corpus. Fourth, our feature set achieved 79.052 on macro F-score, which has a 7.59% improvement over Baseline 2 and a 2.41% improvement over the feature set of (Jochim & Schütze, 2012). Compared to the previously reported results on the same dataset, our method improved more than 20%(66% in (Dong & Schäfer, 2011) and 60.7% in (Jochim, 2014)). This may come from the effectiveness of new features. Table 1 shows the more detailed performance comparison between our feature set and (Jochim & Schütze, 2012). We can see that our feature set is superior to (Jochim & Schütze, 2012) on most indicators.

4 Conclusion

In this poster we address the problem of citation function classification, which could be the key component for the next generation citation analysis and significant technique for constructing intellectual digital library. In order to improve the citation classification performance, we propose new lexical and syntactic features by mining unique linguistic patterns in citation context. A complete comparison experiment is conducted and results show the effectiveness of our features.

In the future, we would like to test the robustness of our model by experimenting on a large-scale corpus and look for better method to improve the model performance. In addition, seeking for potential applications would be another meaningful direction.

	Jochim and Schütze (2012)				Our			
	Correct	Precision	Recall	Macro-F	Correct	Precision	Recall	Macro-F
Idea	**98**	79.67%	**75.97%**	77.78%	97	**84.35%** (**+5.87**)	75.19% (-1.03)	**79.51%** (**+2.08**)
Basis	315	**78.75%**	74.29%	76.46%	**318**	78.52% (-0.29)	**75%** (**+0.96**)	**76.72%** (**+0.34**)
Comparison	39	73.58%	55.71%	63.41%	**41**	**82%** (**+11.44**)	**58.57%** (**+5.13**)	**68.33%** (**+7.76**)
Background	1072	89.33%	92.97%	91.12%	**1081**	**89.64%** (**+0.31**)	**93.76%** (**+0.85**)	**91.65%** (**+0.58**)
Total	1524	80.33%	74.74%	77.19%	1537	**83.63%** (**+4.11**)	**75.63%** (**+1.19**)	**79.05%** (**+2.41**)

Table 1. Detailed performance comparison between feature set in (Jochim & Schütze, 2012) and ours

5 References

Abu-Jbara, A., & Radev, D. (2012). *Reference scope identification in citing sentences.* Paper presented at the Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Athar, A. (2011). *Sentiment analysis of citations using sentence structure-based features.* Paper presented at the Proceedings of the ACL 2011 student session.

Dong, C., & Schäfer, U. (2011). *Ensemble-style Self-training on Citation Classification.* Paper presented at the IJCNLP.

Garzone, M. A. (1997). *Automated classification of citations using linguistic semantic grammars.* The University of Western Ontario.

Jochim, C. (2014). *Natural language processing and information retrieval methods for intellectual property analysis.* Universitätsbibliothek der Universität Stuttgart, Stuttgart. Retrieved from http://elib.uni-stuttgart.de/opus/volltexte/2014/9634

Jochim, C., & Schütze, H. (2012). Towards a generic and flexible citation classifier based on a faceted classification scheme.

Li, X., He, Y., Meyers, A., & Grishman, R. (2013). *Towards Fine-grained Citation Function Classification.* Paper presented at the RANLP.

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., & Lin, Y. (2013). Literature retrieval based on citation context. *Scientometrics*, 1-15.

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9): Springer.

Qazvinian, V., & Radev, D. R. (2008). *Scientific paper summarization using citation summary networks.* Paper presented at the Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1.

Radoulov, R. (2008). Exploring automatic citation classification.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). *Automatic classification of citation function.* Paper presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.

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

6 Table of Figures

7 Table of Tables