Automatic Classification of Citation Function by New Linguistic Features

Rui Meng[1], Wei Lu[2], Yu Chi[1] and Shuguang Han[1]
[1]School of Information Sciences, University of Pittsburgh, USA
[2]School of Information Management, Wuhan University, China

Abstract
Citation function presents the functional role of a reference in its citing article. These functional information enrich the citation analysis in a semantic perspective and can be used for improving the applications of citation analysis. Though many works on automatic classification have been done, the performance of existing studies cannot satisfy the requirement of analysis on large-scale academic data. In order to overcome the performance bottleneck, in this poster we present some useful features by analyzing and finding unique linguistic patterns in citation context. Our experiments on existing dataset shows the effectiveness of these new features with Support Vector Machine. The performance reaches 86.54% accuracy and a macro F-score of 0.795, which gains an improvement over 20% than previous study on the same dataset.

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: rui.meng@pitt.edu, weilu@whu.edu.cn, shh69@pitt.edu and yuc73@pitt.edu

1 Introduction

Traditional citation analysis, which is merely based on the citation network, has been blamed for simplifying the authors' real motivation of citing to a linear and equal relationship. The citation content contains abundant semantic information, which could be used to enrich the presentation of classic citation network 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 attracted to study recognizing citation's natures automatically, which include citation sentiment (Athar, 2011), citation function (Teufel, Siddharthan, & Tidhar, 2006) and citation importance (Wan & Liu, 2014). Among them, citation function is considered to be the most important nature as it presents different role of citation in scientific literature, from introducing related research background to acknowledging the important ideas used in current paper.

A lot of efforts towards a fine-grained citation function classification have been done. Many classification schemes have been introduced by early researchers by manually analyze hundreds of papers. These schemes differ from each other mainly in the granularity and the research domain they concern about. In order to overcome the limitation of manually annotation, researchers made attempt to annotate the citation function automatically. Garzone (Garzone, 1997) built a rule-based classifier based on his own annotation scheme which contains 35 categories. Teufel et al.(Teufel et al., 2006) is the first one proposed to use machine learning method to classify citation function. They trained a classifier by using the IBk algorithm based on a modified classification scheme containing 12 categories and reached a fairly good performance (0.57 of Macro-F). Radoulov (Radoulov, 2008) framed the problem of classifying citations as a word disambiguation task. An improved classification scheme reduced the 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 et al. (Dong & Schäfer, 2011) designed feature sets from the aspects of textual, physical and syntactic, and used a semi-supervised algorithm to make use of unlabeled data, which is meaningful for a small training datasets. Charles et al. (Jochim & Schütze, 2012) gave a systematic investigation of features used in previous research and introduced their new features which show strong improvement.

So far the most significant barriers prevents automatic citation classification from real application is its poor performance. For example Dong et al. (Dong & Schäfer, 2011) achieved 0.67 of macro F-score on 4 categories scheme, other studies based on more categories classification scheme mostly performed worse. In order to overcome the bottleneck of performance on citation classification, some useful features are introduced to reveal the function of certain citation. Also a more powerful classifier Support Vector

Machine is used in our experiment. Our experimental results on existing dataset shows the effectiveness of new features. The performance reaches 86.54% accuracy and a macro F-score of 0.795, which gains an improvement over 20% than previous study on the same dataset.

2 Methods

2.1 Corpus and Classification Scheme

Most previous works conducted automatic classification research based on their own proposed classification scheme and self-annotated corpus. Though we have proposed a fine-grained classification scheme in preliminary research (Lu, Meng, & Liu, 2014), we decide to use the corpus and scheme of previous work in order to make comparison with previous research. Dong et al. (Dong & Schäfer, 2011) designed a four categories classification scheme, which covers most general citation functions and can be extended to a more fine-grained categories easily. The four categories are as follows:

- Background: citing in order to describe the research background of current work.
- Fundamental idea (Idea for short): previous work inspired or gave specific hints on the current work.
- Technical basis (Basis for short): important tools, methods, data and other resources used or adapted in the current work.
- Comparison: current work compares methods or results with the cited work.

Dong et al. annotated 1768 instances of citation function which extracted from 122 conference papers of ACL Anthology. And the number of instances in each category is:

Background : Idea : Basis : Comparison = 1150 : 421 : 127 : 70

2.2 Description of Features

We investigated most of feature set used in previous studies in order to find omitted significant features and compare with ourselves latter. We list all the features used in our work as follows.

- Word-level features. First part of word-level feature is the n-grams introduced by (Athar, 2011). In this work we use unigram only as it's robust enough to capture key lexical information without introduce too much noises. Another main part of word-level feature is the cue words used by Dong et al. (Dong & Schäfer, 2011), especially the subject cue can distinguish the informative categories from uninformative ones significantly. Other word-level features includes modality words, main verb and root verb introduced by (Jochim & Schütze, 2012). Besides, we find that digits and percentages occur frequently in Comparison category, and words denote future work commonly occur in Background category. Thus we added two Boolean features to see whether the citing sentence contains these words.
- Syntactic features. The most common syntactic feature is the dependency relations which showed notable improvement by (Athar, 2011). (Dong & Schäfer, 2011) used regular expression to capture seven types of syntactic patterns. (Jochim & Schütze, 2012) designed several detail feature like whether the citation is labeled as a constituent in sentence, whether pronoun is linked to a comparative. (Abu-Jbara & Radev, 2012) and (Li, He, Meyers, & Grishman, 2013) extracted the signal words which linked to citation marks. Inspired by the subject cue feature from (Dong & Schäfer, 2011), we find that the verbs and adjectives linked to subject cue by dependency relation play significant roles in recognizing citation function. Thus we extract these linked words and relations out as important features.

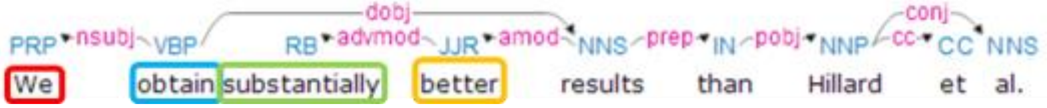

Figure 1. Example shows the key verb and adjective connected to first pronoun

- Physical features. This feature set contains the location and frequency information of each citation. (Dong & Schäfer, 2011) mapped the citation located section into six predefined categories (Introduction, Related work, Method, Experiment, Evaluation and Conclusion) as the location of this citation. Also the number of other citations in the citation sentence as well as in its context is an effective feature to see the importance of this citation.
- Other features. Self-citing feature is first introduced by (Teufel et al., 2006) which assumes that self-citing may indicate important citing relation. Named-entity recognition is used to find whether

the citation is related to resource or tool, which is considered to be a useful feature in recognizing certain categories. We extracted this feature by cue words instead of building a NER tagger.

3 Experimental Results

In this section we introduce our experimental setup and discuss the result of our method against the methods used in previous research. Two baseline experiments are conducted, one only includes the unigram and the other includes both unigram and dependency relation. The purpose of baseline experiments is to see what the performance would be on simplest text features. Our feature set includes all the features discussed in section 2.2, and we also reimplement the feature sets proposed by (Dong & Schäfer, 2011), (Abu-Jbara & Radev, 2012) and (Jochim & Schütze, 2012) as comparison. All the experiments are conducted on Support Vector Machine classifier with RBF kernel on optimal parameters and in a 10-fold cross validation. Also a feature selection based on information gain is conducted in order to evaluate the effectiveness of different features.

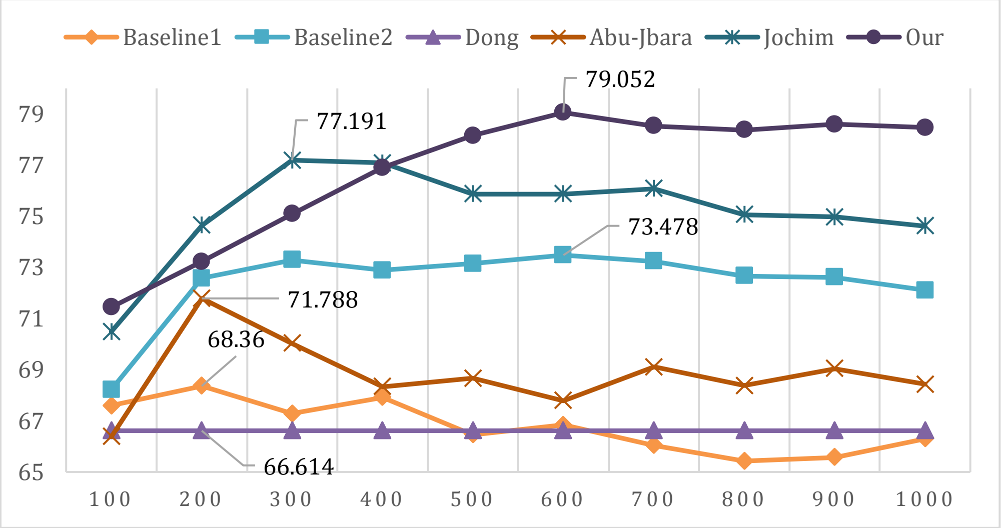

Figure 2. Macro F-score of each feature set on different size of feature selection

Macro F-score is regarded as the main measure instead of accuracy is due to the severely skewed classes. Figure 2 shows the classification performance of each feature set. The first we can see is only small number of features are useful for this task. More than 30,000 features are generated in our feature set, but it performs best on only 600 features. Secondly, two baseline experiments performed pretty good. This tells us simple text features can achieve a fairly good performance on citation classification, but further improvement is difficult. 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 complexity of real situation.

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 results from the effectiveness of new features and more powerful classifier. Table 1 shows the more detailed performance comparison between our feature set and (Jochim & Schütze, 2012) and our feature set is superior to Jochim's on most indicators.

4 Conclusion

In this poster we address the problem of citation function classification, which could be the key to next generation of citation analysis and significant technique for constructing intellectual digital library. In order to overcome the performance bottleneck of citation classification, we proposed new lexical and syntactic features by analyzing and finding unique linguistic patterns in citation context. A complete comparison experiment is conducted and results show the effectiveness of our features with Support Vector Machine. The performance reaches a macro F-score of 0.795, which gains an improvement over 20% than previous study on the same dataset.

In future work, we would like to test the robustness of our model by experiment on large-scale corpus and look for better method to improve the model performance. In addition, exploring the potential application would be a meaningful effort to take.

	Jochim				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 comparison between Jochim's and our feature set

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.

Jiang, Z., Liu, X., & Chen, Y. (2015). Recovering uncaptured citations in a scholarly network: A two-step citation analysis to estimate publication importance. *Journal of the Association for Information Science and Technology*.

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.

Lu, W., Meng, R., & Liu, X. (2014). A Deep Scientific Literature Mining-Oriented Framework for Citation Content Annotation. *Journal of Library Science in China, 40*(6), 12. doi: 10.13530/j.cnki.jlis.140029

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*.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490-1503.