Context-Enhanced Citation Sentiment Detection

Awais Athar
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
Cambridge, CB3 0FD, U.K.
awais.athar@cl.cam.ac.uk

Simone Teufel
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
Cambridge, CB3 0FD, U.K.
simone.teufel@cl.cam.ac.uk

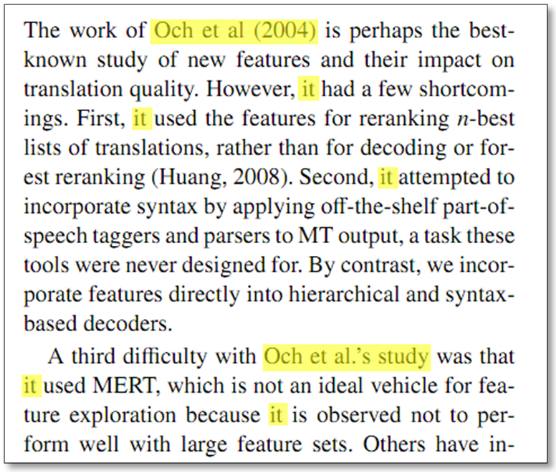
Abstract

Sentiment analysis of citations in scientific papers and articles is a new and interesting problem which can open up many exciting new applications in bibliographic search and bibliometrics. Current work on citation sentiment detection focuses on only the citation sentence. In this paper, we address the problem of context-enhanced citation sentiment detection. We present a new citation sentiment corpus which has been annotated to take the dominant sentiment in the entire citation context into account. We believe that this gold standard is closer to the truth than annotation that looks only at the citation sentence itself. We then explore the effect of context windows of different lengths on the performance of a state-of-the-art citation sentiment detection system when using this context-enhanced gold standard definition.

1 Introduction

Sentiment analysis of citations in scientific papers and articles is a new and interesting problem. It can open up many exciting new applications in bibliographic search and in bibliometrics, i.e., the automatic evaluation of the influence and impact of individuals and journals via citations. Automatic detection of citation sentiment can also be used as a first step to scientific summarisation (Abu-Jbara and Radev, 2011). Alternatively, it can help researchers during search, e.g., by identifying problems with a particular approach, or by helping to recognise unaddressed issues and possible gaps in the current research.

However, there is a problem with the expression of sentiment in scientific text. Conventionally, the writing style in scientific writing is meant to be objective. Any personal bias by authors has to be hedged (Hyland, 1995). Negative sentiment is politically particularly dangerous (Ziman, 1968), and some authors have documented the strategy of prefacing the intended criticism by slightly disingenuous praise (MacRoberts and MacRoberts, 1984). This makes the problem of identifying such opinions particularly challenging. This non-local expression of sentiment has been observed in other genres as well (Wilson et al., 2009; Polanyi and Zaenen, 2006).



The work of Och et al (2004) is perhaps the best-known study of new features and their impact on translation quality. However, it had a few shortcomings. First, it used the features for reranking *n*-best lists of translations, rather than for decoding or forest reranking (Huang, 2008). Second, it attempted to incorporate syntax by applying off-the-shelf part-of-speech taggers and parsers to MT output, a task these tools were never designed for. By contrast, we incorporate features directly into hierarchical and syntax-based decoders.

A third difficulty with Och et al.'s study was that it used MERT, which is not an ideal vehicle for feature exploration because it is observed not to perform well with large feature sets. Others have in-

Figure 1: Example of anaphora in citations

A typical case is illustrated in Figure 1. While the first sentence praises some aspects of the cited paper, the remaining sentences list its shortcomings. It is clear that criticism is the intended sentiment, but

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 597–601,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

if we define our gold standard only by looking at the citation sentence, we lose a significant amount of sentiment hidden in the text. Given that most citations are neutral (Spiegel-Rosing, 1977; Teufel et al., 2006), this makes it ever more important to recover what explicit sentiment there is from the context of the citation.

However, the dominant assumption in current citation identification methods (Ritchie et al., 2008; Radev et al., 2009) is that the sentiment present in the citation sentence represents the true sentiment of the author towards the cited paper. This is due to the difficulty of determining the relevant context, whereas it is substantially easier to identify the citation sentence. In our example above, however, such an approach would lead to the wrong prediction of praise or neutral sentiment.

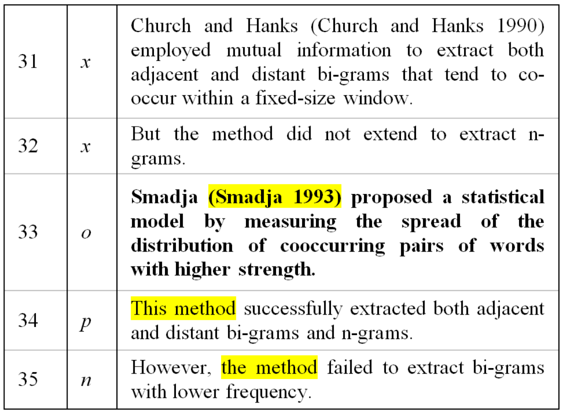
In this paper, we address the problem of context-enhanced citation sentiment detection. We present a new citation sentiment corpus where each citation has been annotated according to the dominant sentiment in the corresponding citation context. We claim that this corpus is closer to the truth than annotation that considers only the citation sentence itself. We show that it increases citation sentiment coverage, particularly for negative sentiment. Using this gold standard, we explore the effect of assuming context windows of different but fixed lengths on the performance of a state-of-the-art citation sentiment detection system where the sentiment of citation is considered in the entire context of the citation and more than one single sentiment can be assigned. Previous approaches neither detect citation sentiment and context simultaneously nor use as large a corpus as we do.

2 Corpus Construction

We chose the dataset used by Athar (2011) comprising 310 papers taken from the ACL Anthology (Bird et al., 2008). The citation summary data from the ACL Anthology Network[1] (Radev et al., 2009) was used. This dataset is rather large (8736 citations) and since manual annotation of context for each citation is a time consuming task, a subset of 20 papers were selected corresponding to approximately 20% of the original dataset.

We selected a four-class scheme for annotation. Every sentence that is in a window of 4 sentences of the citation and does not contain any direct or indirect mention of the citation was labelled as being excluded (x). The window length was motivated by recent research (Qazvinian and Radev, 2010) which shows the best score for a four-sentence boundary when detecting non-explicit citation. The rest of the sentences were marked either positive (p), negative (n) or objective/neutral (o).

A total of 1,741 citations were annotated. Although this annotation was performed by the first author only, we know from previous work that similar styles of annotation can achieve acceptable inter-annotator agreement (Teufel et al., 2006). An example annotation for Smadja (1993) is given in Figure 2, where the first column shows the line number and the second one shows the class label.



31	x	Church and Hanks (Church and Hanks 1990) employed mutual information to extract both adjacent and distant bi-grams that tend to co-occur within a fixed-size window.
32	x	But the method did not extend to extract n-grams.
33	o	**Smadja (Smadja 1993) proposed a statistical model by measuring the spread of the distribution of cooccurring pairs of words with higher strength.**
34	p	This method successfully extracted both adjacent and distant bi-grams and n-grams.
35	n	However, the method failed to extract bi-grams with lower frequency.

Figure 2: Example annotation of a citation context.

To compare our work with Athar (2011), we also applied a three-class annotation scheme. In this method of annotation, we merge the citation context into a single sentence. Since the context introduces more than one sentiment per citation, we marked the citation sentiment with the last sentiment mentioned in the context window as this is pragmatically most likely to be the real intention (MacRoberts and MacRoberts, 1984).

As is evident from Table 1, including the 4 sentence window around the citation more than doubles the instances of subjective sentiment, and in the case of *negative* sentiment, this proportion rises to 3. In light of the overall sparsity of detectable citation sentiment in a paper, and of the envisaged applica-

[1] http://www.aclweb.org

tions, this is a very positive result. The reason for this effect is most likely "sweetened criticism" – authors' strategic behaviour of softening the effect of criticism among their peers (Hornsey et al., 2008).

	Without Context	With Context
o	87%	73%
n	5%	17%
p	8%	11%

Table 1: Distribution of classes.

3 Experiments and Results

We represent each citation as a feature set in a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) framework and use n-grams of length 1 to 3 as well as dependency triplets as features. The dependency triplets are constructed by merging the relation, governor and dependent in a single string, for instance, the relation *nsubj*(*failed*, *method*) is represented as `nsubj_failed_method`. This setup has been shown to produce good results earlier as well (Pang et al., 2002; Athar, 2011).

The first set of experiments focuses on simultaneous detection of sentiment and context sentences. For this purpose, we use the four-class annotated corpus described earlier. While the original annotations were performed for a window of length 4, we also experiment with asymmetrical windows of l sentences preceding the citation and r sentences succeeding it. The detailed results are given in Table 2.

l	r	x	o	n	p	F_{macro}	F_{micro}
0	**0**	**-**	**1509**	**86**	**146**	**0.768**	**0.932**
1	1	2823	1982	216	200	0.737	0.820
2	2	5984	2214	273	218	0.709	0.851
3	3	9170	2425	318	234	0.672	0.875
4	4	12385	2605	352	252	0.680	0.892
0	4	5963	2171	322	215	0.712	0.853
0	3	4380	2070	293	201	0.702	0.832
0	2	2817	1945	258	193	0.701	0.801
0	1	1280	1812	206	182	0.717	0.777

Table 2: Results for joint context and sentiment detection.

Because of the skewed class distribution, we use both the F_{macro} and F_{micro} scores with 10-fold cross-validation. The baseline score, shown in bold, is obtained with no context window and is comparable to the results reported by Athar (2011). However, we can observe that the F scores decrease as more context is introduced. This may be attributed to the increase in the vocabulary size of the n-grams and a consequent reduction in the discriminating power of the decision boundaries. These results show that the task of jointly detecting sentiment and context is a hard problem.

For our second set of experiments, we use the three-class annotation scheme. We merge the text of the sentences in the context windows as well as their dependency triplets to obtain the features. The results are reported in Table 3 with best results in bold. Although these results are not better than the context-less baseline, the reason might be data sparsity since existing work on citation sentiment analysis uses more data (Athar, 2011).

l	r	F_{macro}	F_{micro}
1	1	0.638	0.827
2	2	0.620	0.793
3	3	0.629	0.786
4	4	0.628	0.771
0	4	0.643	0.796
0	3	0.658	0.816
0	2	0.642	0.824
0	1	**0.731**	**0.871**

Table 3: Results using different context windows.

4 Related Work

While different schemes have been proposed for annotating citations according to their function (Spiegel-Rosing, 1977; Nanba and Okumura, 1999; Garzone and Mercer, 2000), the only recent work on citation sentiment detection using a relatively large corpus is by Athar (2011). However, this work does not handle citation context. Piao et al. (2007) proposed a system to attach sentiment information to the citation links between biomedical papers by using existing semantic lexical resources.

A common approach for sentiment detection is to use a labelled lexicon to score sentences (Hatzivassiloglou and McKeown, 1997; Turney, 2002; Yu and Hatzivassiloglou, 2003). However, such approaches have been found to be highly topic dependent (Engström, 2004; Gamon and Aue, 2005; Blitzer et al., 2007).

Teufel et al. (2006) worked on a 2,829 sentence citation corpus using a 12-class classification scheme. Although they used context in their annotation, their focus was on determining the author's reason for citing a given paper. This task differs from citation sentiment, which is in a sense a "lower level" of analysis.

For implicit citation extraction, Kaplan et al. (2009) explore co-reference chains for citation extraction using a combination of co-reference resolution techniques. However, their corpus consists of only 94 sentences of citations to 4 papers which is likely to be too small to be representative. The most relevant work is by Qazvinian and Radev (2010) who extract only the non-explicit citations for a given paper. They model each sentence as a node in a graph and experiment with various window boundaries to create edges between neighbouring nodes. However, their dataset consists of only 10 papers and their annotation scheme differs from our four-class annotation as they do not deal with any sentiment.

5 Conclusion

In this paper, we focus on automatic detection of citation sentiment using the citation context. We present a new corpus and show that ignoring the citation context would result in loss of a lot of sentiment, specially criticism towards the cited paper. We also report the results of the state-of-the-art citation sentiment detection systems on this corpus when using this context-enhanced gold standard definition.

Future work directions may include improving the detection algorithms by filtering the context sentences more intelligently. For this purpose, existing work on coreference resolution (Lee et al., 2011) may prove to be useful. Context features may also be used for first filtering citations which have been mentioned only in passing, and then applying context based sentiment classification to the remaining significant citations.

References

A. Abu-Jbara and D. Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proc. of ACL*.

A. Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proc of ACL*, page 81.

S. Bird, R. Dale, B.J. Dorr, B. Gibson, M.T. Joseph, M.Y. Kan, D. Lee, B. Powley, D.R. Radev, and Y.F. Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. of LREC*.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. of ACL*, number 1.

C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

C. Engström. 2004. Topic dependence in sentiment classification. University of Cambridge.

M. Gamon and A. Aue. 2005. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proc. of the ACL*.

M. Garzone and R. Mercer. 2000. Towards an automated citation classifier. *Advances in Artificial Intelligence*.

V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proc. of ACL*, page 181.

M.J. Hornsey, E. Robson, J. Smith, S. Esposo, and R.M. Sutton. 2008. Sugaring the pill: Assessing rhetorical strategies designed to minimize defensive reactions to group criticism. *Human Communication Research*, 34(1):70–98.

K. Hyland. 1995. The Author in the Text: Hedging Scientific Writing. *Hong Kong papers in linguistics and language teaching*, 18:11.

D. Kaplan, R. Iida, and T. Tokunaga. 2009. Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In *Proc. of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. *ACL HLT 2011*.

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

H. Nanba and M. Okumura. 1999. Towards multi-paper summarization using reference information. In *IJCAI*, volume 16, pages 926–931. Citeseer.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. of EMNLP*.

S. Piao, S. Ananiadou, Y. Tsuruoka, Y. Sasaki, and J. McNaught. 2007. Mining opinion polarity relations of citations. In *International Workshop on Computational Semantics (IWCS)*. Citeseer.

L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, pages 1–10.

V. Qazvinian and D.R. Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proc. of ACL*.

D.R. Radev, M.T. Joseph, B. Gibson, and P. Muthukrishnan. 2009. A Bibliometric and Network Analysis of the field of Computational Linguistics. *Journal of the American Soc. for Info. Sci. and Tech.*

A. Ritchie, S. Robertson, and S. Teufel. 2008. Comparing citation contexts for information retrieval. In *Proc. of ACM conference on Information and knowledge management*, pages 213–222. ACM.

I. Spiegel-Rosing. 1977. Science studies: Bibliometric and content analysis. *Social Studies of Science*.

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

P.D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Comp. Ling.*, 35(3):399–433.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. of EMNLP*, page 136.

J.M. Ziman. 1968. *Public Knowledge: An essay concerning the social dimension of science*. Cambridge Univ. Press, College Station, Texas.