Sentiment Analysis of Citations using Sentence Structure-Based Features

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

Abstract

Sentiment analysis of citations in scientific papers and articles is a new and interesting problem due to the many linguistic differences between scientific texts and other genres. In this paper, we focus on the problem of automatic identification of positive and negative sentiment polarity in citations to scientific papers. Using a newly constructed annotated citation sentiment corpus, we explore the effectiveness of existing and novel features, including n-grams, specialised science-specific lexical features, dependency relations, sentence splitting and negation features. Our results show that 3-grams and dependencies perform best in this task; they outperform the sentence splitting, science lexicon and negation based features.

1 Introduction

Sentiment analysis is the task of identifying positive and negative opinions, sentiments, emotions and attitudes expressed in text. Although there has been in the past few years a growing interest in this field for different text genres such as newspaper text, reviews and narrative text, relatively less emphasis has been placed on extraction of opinions from scientific literature, more specifically, citations. Analysis of citation sentiment would open up many exciting new applications in bibliographic search and in bibliometrics, i.e., the automatic evaluation the influence and impact of individuals and journals via citations.

Existing bibliometric measures like H-Index (Hirsch, 2005) and adapted graph ranking algorithms like PageRank (Radev et al., 2009) treat all citations as equal. However, Bonzi (1982) argued that if a cited work is criticised, it should consequently carry lower or even negative weight for bibliometric measures. Automatic citation sentiment detection is a prerequisite for such a treatment.

Moreover, citation sentiment detection can also help researchers during search, by detecting problems with a particular approach. It can be used as a first step to scientific summarisation, enable users to recognise unaddressed issues and possible gaps in the current research, and thus help them set their research directions.

For other genres a rich literature on sentiment detection exists and researchers have used a number of features such as n-grams, presence of adjectives, adverbs and other parts-of-speech (POS), negation, grammatical and dependency relations as well as specialised lexicons in order to detect sentiments from phrases, words, sentences and documents. State-of-the-art systems report around 85-90% accuracy for different genres of text (Nakagawa et al., 2010; Yessenalina et al., 2010; Täckström and McDonald, 2011).

Given such good results, one might think that a sentence-based sentiment detection system trained on a different genre could be used equally well to classify citations. We argue that this might not be the case; our citation sentiment recogniser uses specialised training data and tests the performance of specialised features against current state-of-the-art features. The reasons for this are based on the following observations:

- Sentiment in citations is often hidden. This might

Proceedings of the ACL-HLT 2011 Student Session, pages 81–87,
Portland, OR, USA 19-24 June 2011. ©2011 Association for Computational Linguistics

be because of the general strategy to avoid overt criticism due to the sociological aspect of citing (MacRoberts and MacRoberts, 1984; Thompson and Yiyun, 1991). Ziman (1968) states that many works are cited out of "politeness, policy or piety". Negative sentiment, while still present and detectable for humans, is expressed in subtle ways and might be hedged, especially when it cannot be quantitatively justified (Hyland, 1995).

While SCL has been successfully applied to POS tagging and Sentiment Analysis (Blitzer et al., 2006), its effectiveness for parsing was ***rather unexplored****.*

- Citation sentences are often neutral with respect to sentiment, either because they describe an algorithm, approach or methodology objectively, or because they are used to support a fact or statement.

  There are five different IBM translation models (Brown et al. , 1993).

  This gives rise to a far higher proportion of objective sentences than in other genres.

- Negative polarity is often expressed in contrastive terms, e.g. in evaluation sections. Although the sentiment is indirect in these cases, its negativity is implied by the fact that the authors' own work is clearly evaluated positively in comparison.

  This method was shown to ***outperform*** *the class based model proposed in (Brown et al., 1992) . . .*

- There is also much variation between scientific texts and other genres concerning the lexical items chosen to convey sentiment. Sentiment carrying science-specific terms exist and are relatively frequent, which motivates the use of a sentiment lexicon specialised to science.

  Similarity-based smoothing (Dagan, Lee, and Pereira 1999) provides an ***intuitively appealing*** *approach to language modeling.*

- Technical terms play a large role overall in scientific text (Justeson and Katz, 1995). Some of these carry sentiment as well.

  Current ***state of the art*** *machine translation systems (Och, 2003) use phrasal (n-gram) features . . .*

  For this reason, using higher order n-grams might prove to be useful in sentiment detection.

- The scope of influence of citations varies widely from a single clause (as in the example below) to several paragraphs:

  *As reported in Table 3, small increases in METEOR (****Banerjee and Lavie, 2005****), BLEU (Papineni et al., 2002) and NIST scores (Doddington, 2002) suggest that . . .*

  This affects lexical features directly since there could be "sentiment overlap" associated with neighbouring citations. Ritchie et al. (2008) showed that assuming larger citation scopes has a positive effect in retrieval. We will test the opposite direction here, i.e., we assume short scopes and use a parser to split sentences, so that the features associated with the clauses not directly connected to the citation are disregarded.

We created a new sentiment-annotated corpus of scientific text in the form of a sentence-based collection of over 8700 citations. Our experiments use a supervised classifier with the state-of-the-art features from the literature, as well as new features based on the observations above. Our results show that the most successful feature combination includes dependency features and n-grams longer than for other genres ($n = 3$), but the assumption of a smaller scope (sentence splitting) decreased results.

2 Training and Test Corpus

We manually annotated 8736 citations from 310 research 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. We identified the actual text of the citations by regular expressions and replaced it with a special token <*CIT*> in order to remove any lexical bias associated with proper names of researchers. We labelled each sentence as positive, negative or objective, and separated 1472 citations for development and training. The rest were used as the test set containing 244 negative, 743 positive and 6277 objective citations. Thus our dataset is heavily skewed, with subjective citations accounting for only around 14% of the corpus.

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

3 Features

We represent each citation as a feature set in a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) framework which has been shown to produce good results for sentiment classification (Pang et al., 2002). The corpus is processed using WEKA (Hall et al., 2008) and the Weka LibSVM library (EL-Manzalawy and Honavar, 2005; Chang and Lin, 2001) with the following features.

3.1 Word Level Features

In accordance with Pang et al. (2002), we use unigrams and bigrams as features and also add 3-grams as new features to capture longer technical terms. POS tags are also included using two approaches: attaching the tag to the word by a delimiter, and appending all tags at the end of the sentence. This may help in distinguishing between homonyms with different POS tags and signalling the presence of adjectives (e.g., JJ) respectively. Name of the primary author of the cited paper is also used as a feature.

A science-specific sentiment lexicon is also added to the feature set. This lexicon consists of 83 polar phrases which have been manually extracted from the development set of 736 citations. Some of the most frequently occurring polar phrases in this set consists of adjectives such as *efficient*, *popular*, *successful*, *state-of-the-art* and *effective*.

3.2 Contextual Polarity Features

Features previously found to be useful for detecting phrase-level contextual polarity (Wilson et al., 2009) are also included. Since the task at hand is sentence-based, we use only the sentence-based features from the literature e.g., presence of subjectivity clues which have been compiled from several sources[2] along with the number of adjectives, adverbs, pronouns, modals and cardinals.

To handle negation, we include the count of negation phrases found within the citation sentence. Similarly, the number of valance shifters (Polanyi and Zaenen, 2006) in the sentence are also used. The polarity shifter and negation phrase lists have been taken from the OpinionFinder system (Wilson et al., 2005).

3.3 Sentence Structure Based Features

We explore three different feature sets which focus on the lexical and grammatical structure of a sentence and have not been explored previously for the task of sentiment analysis of scientific text.

3.3.1 Dependency Structures

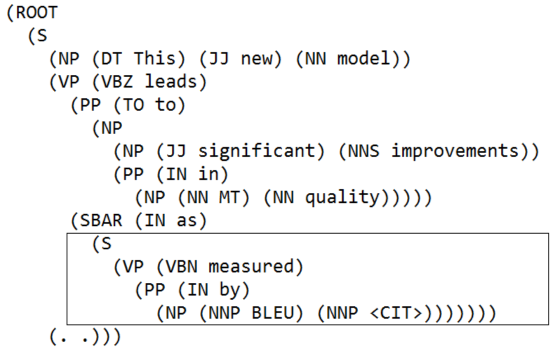
The first set of these features include typed dependency structures (de Marneffe and Manning, 2008) which describe the grammatical relationships between words. We aim to capture the long distance relationships between words. For instance in the sentence below, the relationship between *results* and *competitive* will be missed by trigrams but the dependency representation captures it in a single feature `nsubj_competitive_results`.

<CIT> showed that the results for French-English were competitive to state-of-the-art alignment systems.

A variation we experimented with, but gave up on as it did not show any improvements, concerns backing-off the dependent and governor to their POS tags (Joshi and Penstein-Rosé, 2009).

3.3.2 Sentence Splitting

Removing irrelevant polar phrases around a citation might improve results. For this purpose, we split each sentence by trimming its parse tree. Walking from the citation node (<*CIT*>) towards the root, we select the subtree rooted at the first sentence node (*S*) and ignore the rest. For example, in Figure 1, the cited paper is not included in the scope of the discarded polar phrase *significant improvements*.



```
(ROOT
  (S
    (NP (DT This) (JJ new) (NN model))
    (VP (VBZ leads)
      (PP (TO to)
        (NP
          (NP (JJ significant) (NNS improvements))
          (PP (IN in)
            (NP (NN MT) (NN quality)))))
      (SBAR (IN as)
        (S
          (VP (VBN measured)
            (PP (IN by)
              (NP (NNP BLEU) (NNP <CIT>)))))))
    (. .)))
```

Figure 1: An example of parse tree trimming

[2] Available for download at http://www.cs.pitt.edu/mpqa/

3.3.3 Negation

Dependencies and parse trees attach negation nodes, such as *not*, to the clause subtree and this shows no interaction with other nodes with respect to valence shifting. To handle this effect, we take a simple window-based inversion approach. All words inside a k-word window of any negation term are suffixed with a token *_neg* to distinguish them from their non-polar versions. For example, a 2-word negation window inverts the polarity of the positive phrase *work well* in the sentence below.

Turney's method did not work_neg well_neg although they reported 80% accuracy in <CIT>.

The negation term list has been taken from the OpinionFinder system. Khan (2007) has shown that this approach produces results comparable to grammatical relations based negation models.

4 Results

Because of our skewed dataset, we report both the macro-F and the micro-F scores using 10-fold cross-validation (Lewis, 1991). The bold values in Table 1 show the best results.

Features	**macro-F**	**micro-F**
1 grams	0.581	0.863
1-2 grams	0.592	0.864
1-3 grams	0.597	0.862
″ + POS	0.535	0.859
″ + POS (tokenised)	0.596	0.859
″ + scilex	0.597	0.860
″ + wlev	0.535	0.859
″ + cpol	0.418	0.859
″ + dep	0.760	0.897
″ + dep + split + neg	0.683	0.872
″ + dep + split	0.642	0.866
″ + dep + neg	**0.764**	**0.898**

Table 1: Results using science lexicon (scilex), contextual polarity (cpol), dependencies (dep), negation (neg), sentence splitting (split) and word-level (wlev) features.

The selection of the features is on the basis of improvements over a baseline of 1-3 grams i.e. if a feature (e.g. scilex) did not shown any improvement, it is has been excluded from the subsequent experiments.

The results show that contextual polarity features do not work well on citation text. Adding a science-specific lexicon does not help either. This may indicate that n-grams are sufficient to capture discriminating lexical structures. We find that word level and contextual polarity features are surpassed by dependency features. Sentence splitting does not help, possibly due to longer citation scope. Adding a negation window (k=15) improves the performance but the improvement was not found to be statistically significant. This might be due to skewed class distribution and a larger dataset may prove to be useful.

5 Related Work

While different schemes have been proposed for annotating citations according to their function (Spiegel-Rösing, 1977; Nanba and Okumura, 1999; Garzone and Mercer, 2000), there have been no attempts on citation sentiment detection in a large corpus.

Teufel et al. (2006) worked on a 2829 sentence citation corpus using a 12-class classification scheme. However, this corpus has been annotated for the task of determining the author's reason for citing a given paper and is thus built on top of sentiment of citation. It considers usage, modification and similarity with a cited paper as positive even when there is no sentiment attributed to it. Moreover, contrast between two cited methods (CoCoXY) is categorized as objective in the annotation scheme even if the text indicates that one method performs better than the other. For example, the sentence below talks about a positive attribute but is marked as neutral in the scheme.

Lexical transducers are more efficient for analysis and generation than the classical two-level systems (Koskenniemi,1983) because . . .

Using this corpus is thus more likely to lead to inconsistent representation of sentiment in any system which relies on lexical features. Teufel et al. (2006) group the 12 categories into 3 in an attempt to perform a rough approximation of sentiment analysis over the classifications and report a 0.710 macro-F score. Unfortunately, we have access to only a subset[3] of this citation function corpus. We have extracted 1-3 grams, dependencies and negation features from the reduced citation function dataset and used them in our system with 10-fold cross-validation. This results in an improved macro-F score of 0.797 for the subset. This shows that our system is comparable to Teufel et al. (2006). When this subset is used to test the system trained on our newly annotated corpus, a low macro-F score of 0.484 is achieved. This indicates that there is a mismatch in the annotated class labels. Therefore, we can infer that citation sentiment classification is different from citation function classification.

Other approaches to citation annotation and classification include Wilbur et al. (2006) who annotated a small 101 sentence corpus on focus, polarity, certainty, evidence and directionality. Piao et al. (2007) proposed a system to attach sentiment information to the citation links between biomedical papers.

Different dependency relations have been explored by Dave et al. (2003), Wilson et al. (2004) and Ng et al. (2006) for sentiment detection. Nakagawa et al. (2010) report that using dependencies on conditional random fields with lexicon based polarity reversal results in improvements over n-grams for news and reviews corpora.

A common approach is to use a sentiment labelled lexicon to score sentences (Hatzivassiloglou and McKeown, 1997; Turney, 2002; Yu and Hatzivassiloglou, 2003). Research suggests that creating a general sentiment classifier is a difficult task and existing approaches are highly topic dependent (Engström, 2004; Gamon and Aue, 2005; Blitzer et al., 2007).

6 Conclusion

In this paper, we focus on automatic identification of sentiment polarity in citations. Using a newly constructed annotated citation sentiment corpus, we examine the effectiveness of existing and novel features, including n-grams, scientific lexicon, dependency relations and sentence splitting. Our results show that 3-grams and dependencies perform best in this task; they outperform the scientific lexicon and the sentence splitting features. Future directions include trying to improve the performance by modelling negations using a more sophisticated approach. New techniques for detection of the negation scope such as the one proposed by Councill et al. (2010) might also be helpful in citations. Exploring longer citation scopes by including citation contexts might also improve citation sentiment detection.

[3] This subset contains 591 positive, 59 negative and 1259 objective citations.

References

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 the 6th International Conference on Language Resources and Evaluation Conference (LREC08)*, pages 1755–1759. Citeseer.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 45, page 440.

S. 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.

C.C. Chang and C.J. Lin. 2001. LIBSVM: a library for support vector machines, 2001. *Software available at* `http://www.csie.ntu.edu.tw/cjlin/libsvm`.

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

I.G. Councill, R. McDonald, and L. Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59. Association for Computational Linguistics.

K. Dave, S. Lawrence, and D.M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

M.C. de Marneffe and C.D. Manning. 2008. The Stanford typed dependencies representation. In *COLING*, pages 1–8. Association for Computational Linguistics.

Y. EL-Manzalawy and V. Honavar, 2005. *WLSVM: Integrating LibSVM into Weka Environment*. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

C. Engström. 2004. Topic dependence in sentiment classification. *Unpublished MPhil Dissertation. University of Cambridge*.

M. Gamon and A. Aue. 2005. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 57–64. Association for Computational Linguistics.

M. Garzone and R. Mercer. 2000. Towards an automated citation classifier. *Advances in Artificial Intelligence*, pages 337–346.

D. Hall, D. Jurafsky, and C.D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP*, pages 363–371.

V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of EACL*, pages 174–181. Association for Computational Linguistics.

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

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

M. Joshi and C. Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316. Association for Computational Linguistics.

J.S. Justeson and S.M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(01):9–27.

S. Khan. 2007. *Negation and Antonymy in Sentiment Classification*. Ph.D. thesis, Computer Lab, University of Cambridge.

D.D. Lewis. 1991. Evaluating text categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 312–318.

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

T. Nakagawa, K. Inui, and S. Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In *NAACL HLT*, pages 786–794. Association for Computational Linguistics.

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

V. Ng, S. Dasgupta, and SM Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 611–618. Association for Computational Linguistics.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86. Association for Computational Linguistics.

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)*, pages 366–371. Citeseer.

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

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 Society for Information Science and Technology*, 1001:48109–1092.

A. Ritchie, S. Robertson, and S. Teufel. 2008. Comparing citation contexts for information retrieval. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 213–222. ACM.

I. Spiegel-Rösing. 1977. Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7(1):97–113.

O. Täckström and R. McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the ECIR*.

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

G. Thompson and Y. Yiyun. 1991. Evaluation in the reporting verbs used in academic papers. *Applied linguistics*, 12(4):365.

P.D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics.

W.J. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC bioinformatics*, 7(1):356.

T. Wilson, J. Wiebe, and R. Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *EMNLP*, pages 347–354. Association for Computational Linguistics.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

A. Yessenalina, Y. Yue, and C. Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of EMNLP*, pages 1046–1056, Cambridge, MA, October. Association for Computational Linguistics.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, pages 129–136. Association for Computational Linguistics.

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