"Expresses-an-opinion-about": using corpus statistics in an information extraction approach to opinion mining

**Asad B. Sayeed, Hieu C. Nguyen,
and Timothy J. Meyer**
Department of Computer Science
University of Maryland, College Park
asayeed@cs.umd.edu,
hcnguyen88@gmail.com,
tmeyer1@umd.edu

Amy Weinberg
Institute for Advanced Computer Studies
Department of Linguistics
University of Maryland, College Park
weinberg@umiacs.umd.edu

Abstract

We present a technique for identifying the sources and targets of opinions without actually identifying the opinions themselves. We are able to use an information extraction approach that treats opinion mining as relation mining; we identify instances of a binary "expresses-an-opinion-about" relation. We find that we can classify source-target pairs as belonging to the relation at a performance level significantly higher than two relevant baselines.

This technique is particularly suited to emerging approaches in corpus-based social science which focus on aggregating interactions between sources to determine their effects on socio-economically significant targets. Our application is the analysis of information technology (IT) innovations. This is an example of a more general problem where opinion is expressed using either sub- or supersets of expressive words found in newswire. We present an annotation scheme and an SVM-based technique that uses the local context as well as the corpus-wide frequency of a source-target pair as data to determine membership in "expresses-an-opinion-about". While the presence of conventional subjectivity keywords appears significant in the success of this technique, we are able to find the most domain-relevant keywords without sacrificing recall.

1 Introduction

Two problems in sentiment analysis consist of source attribution and target discovery—who has an opinion, and about what? These problems are usually presented in terms of techniques that relate them to the actual opinion expressed. We have a social science application in which the identification of sources and targets over a large volume of text is more important than identifying the actual opinions particularly in experimenting with social science models of opinion trends. Consequently, we are able to use lightweight techniques to identify sources and targets without using resource-intensive techniques to identify opinionated phrases.

Our application for this work is the discovery of networks of influence among opinion leaders in the IT field. We are interested in answering questions about who the leaders in the field are and how their opinion matches the social and economic success of IT innovation. Consequently, it became necessary for us to construct a system (figure 1) that finds the expressions in text that refer to an opinion leader's activities in promoting or deprecating a technology.

In this paper, we demonstrate an information extraction (Mooney and Bunescu, 2005) approach based in relation mining (Girju et al., 2007) that is effective for this purpose. We describe a technique by which corpus statistics allow us to classify pairs of entities and sentiment analysis targets as instances of an "expresses-an-opinion-about" relation in documents in the IT business press. This genre has the characteristic that many entities and targets are represented within individual sentences and paragraphs. Features based on the

Coling 2010: Poster Volume, pages 1095–1103,
Beijing, August 2010

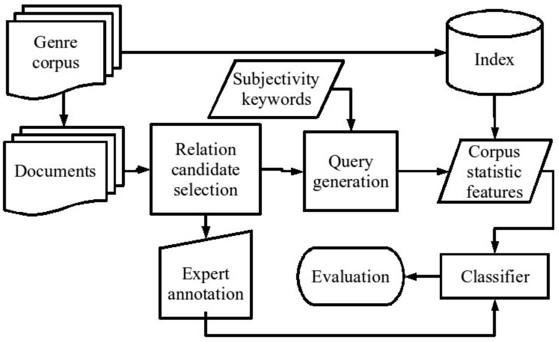

Figure 1: Opinion relation classification system.

frequency counts of query results allow us to train classifiers that allow us to extract "expresses-an-opinion-about" instances, using a very simple annotation strategy to acquire training examples.

In the IT business press, the opinionated language is different from the newswire text for which many extant sentiment tools were developed. We use an existing sentiment lexicon alongside other non-sentiment-specific measures that adapt resources from newswire-developed sentiment analysis projects without imposing the full complexity of those techniques.

1.1 Corpus-based social science

The "expresses-an-opinion-about" relation is a binary relation between opinion sources and targets. Sources include both people—typically known experts, corporate representatives, and other businesspeople—as well as organizations such as corporations and government bodies. The targets are the innovation terms. Therefore, the use of named-entity recognition in this project only focuses on persons and organizations, as the targets are a fixed list.

1.2 Reifying opinion in an application context

A hypothesis implicit in our social science task is that opinion leaders create trends in IT innovation adoption partly by the text that their activities generate in the IT business press. This text has an effect on readers, and these readers act in such a way that in turn may generate more or less prominence for a given innovation—and may also generate further text.

Some of these text-generating activities include expressions of private states in an opinion source (e.g., "I believe that *Web 2.0 is the future*"). These kinds of expressions suggest a particular ontology of opinion analysis involving discourse relations across various types of clauses (Wilson and Wiebe, 2005; Wilson et al., 2005a). However, if we are to track the relative adoption of IT innovations, we must take into account the effect of the text on the reader's opinion about these innovations—there are expressions other than those of private states that have an effect on the reader. These can be considered to be "opinionated acts[1]."

Opinionated acts can include things like purchasing and adoption decisions by organizations. For example:

> And like other top suppliers to **Wal-Mart Stores Inc.**, **BP** has been involved in a mandate to affix *radio frequency identification* tags with embedded electronic product codes to its crates and pallets. (ComputerWorld, January 2005)

In this case, both Wal-Mart and BP have expressed implicit approval for radio frequency identification by adopting it. This may affect the reader's own likelihood of support or adoption of the technology. In this context, we do not directly consider the subjectivity of the opinion source, even though that may be present.

Opinionated acts include things like implications of technology use, not just adoption. We thus define opinion expressions as follows: any expression involving some actor that is likely to affect a reader's own potential to adopt, reject, or speak positively or negatively of a target. This would include "conventional" expressions of private states as well as opinionated acts.

Our definition of "expresses-an-opinion-about" follows immediately. Source A expresses an opinion about target B if an interested third party C's actions towards B may be affected by A's textually recorded actions, in a context where actions

[1] Somasundaran and Wiebe (2009) mention a related category of "pragmatic opinions" that involve world knowledge.

have positive or negative weight (e.g. purchasing, promotion, etc.).

1.3 Domain-specific sentiment detection

We construct a system that uses named-entity recognition and supervised machine learning via SVMs to automatically discover instances of "expresses-an-opinion-about" as a binary relation at reasonably high accuracy and precision.

The advantage of our approach is that, outside of HMM-based named-entity detection (BBN's IdentiFinder), we evade the need for resource-intensive techniques such as sophsticated grammatical models, sequence models, and semantic role labelling (Choi et al., 2006; Kim and Hovy, 2006) by removing the focus on the actual opinion expressed. Then we can use a simple supervised discriminative technique with a joint model of local term frequency information and corpus-wide co-occurrence distributions in order to discover the raw data for opinion trend modelling. The most complex instrument we use from sentiment analysis research on conventional newswire is a sentiment keyword lexicon (Wilson et al., 2005b); furthermore, our techniques allow us to distinguish sentiment keywords that indicate opinion in this domain from keywords that actually indicate that there is no opinion relation between source and target.

While we show that this lightweight technique works well at a paragraph level, it can also be used in conjunction with more resource-intensive techniques used to find "conventional" opinion expressions. Also, the use of topic aspects (Somasundaran and Wiebe, 2009) in conjunction with target names has been associated with an improvement in recall. However, our technique still performs well above the baseline without these improvements.

2 Methodology

2.1 Article preparation

We have a list of IT innovations on which our opinion leader research effort is most closely focused. This list contains common names that refer to these technologies as well as some alternate names and abbreviations. We selected articles at random from the ComputerWorld IT journal that contained mentions of members of the given list. These direct mentions were tagged in the document as XML entities.

Each article was processed by BBN's IdentiFinder 3.3 (Bikel et al., 1999), a named entity recognition (NER) system that tags named mentions of person and organization entities[2].

The articles were then divided into paragraphs. For each paragraph, we generated candidate relations from the entities and innovations mentioned therein. To generate candidates, we paired every entity in the paragraph with every innovation. Redundant pairs are sometimes generated when an entity is mentioned in multiple ways in the paragraph. We eliminated most of these by removing entities whose mentions were substrings of other mentions. For example, "Microsoft" and "Microsoft Corp." are sometimes found in the same paragraph; we eliminate "Microsoft."

2.2 Annotation

We processed 20 documents containing 157 relations in the manner described in the previous section. Then two domain experts (chosen from the authors) annotated every candidate pair in every document according to the following scheme (illustrated in figure 2):

- If the paragraph associated with the candidate pair describes a valid source-target relation, the experts annotated it with `Y`.
- If the paragraph does not actually contain that source-target relation, the experts annotated it with `N`.
- If either the source or the target is misidentified (e.g., errors in named entity recognition), the experts annotated it with `X`.

The Cohen's κ score was 0.6 for two annotators. While this appears to be only moderate agreement, we are still able to achieve good performance in our experiments with this value.

[2] In a separate research effort, we found that IdentiFinder has a high error rate on IT business press documents, so we built a system to reduce the error *post hoc*. We ran this system over the IdentiFinder annotations.

Davis says she has especially enjoyed working with the **PowerPad**'s *bluetooth* interfaces to phones and printers. "It's nice getting into new wireless technology," she says. The *bluetooth* capability will allow couriers to transmit data without docking their devices in their trucks.

Source	Target	Class
Davis	bluetooth	**Y**/N/X
PowerPad	bluetooth	Y/N/**X**

Figure 2: Example paragraph annotation exercise.

We then selected 75 different documents for each annotator and processed and annotated them as above. At this point we have the instances and the classes to which they belong. We labelled 466 instances of Y, 325 instances of N, and 280 instances of X, for a total of 1071 relations.

2.3 Feature vector generation

We have four classes of features for every relation instance. Each type of feature consists of counts extracted from an index of 77,227 ComputerWorld articles from January 1988 to June 2008 generated by the University of Massachusetts search engine Indri (Metzler and Croft, 2004). Each vector is normalized to the unit vector. The index is not stemmed for performance reasons.

The first type of feature consists of simple document frequency statistics for source-target pairs throughout the corpus. The second type consists of document frequency counts of source-target pairs when they are in particularly close proximity to one another. The third type consists of document frequency counts of source target pairs proximate to keywords that reflect subjectivity. The fourth and final type consist of TFIDF scores of vocabulary items in the paragraph containing the putative opinion-holding relation (unigram context features). We use the first three features types to represent the likelihood in the "world" that the source has an opinion about the target and the last feature type to represent the likelihood of the specific paragraph containing an opinion that reflects the source-target relation.

We have a total of 7450 features. Each vector is represented as a sparse array. 806 features represent queries on the Indri index. For all the features, we therefore have 863,226 index queries. We perform the queries in parallel on 25 processors to generate the full feature array, which takes approximately an hour on processors running at 8Ghz. We eliminate all values that are smaller in magnitude than 0.000001 after unit vector normalization.

2.3.1 Frequency statistics

There are two simple frequency statistics features generated from Indri queries. The first is the raw frequency counts of within-document co-occurrences of the source and target in the relation. The second is the mean co-occurrence frequency of the source and target per ComputerWorld document.

2.3.2 Proximity counts

For every relation, we query Indri to check how often the source and the target appear in the same document in the ComputerWorld corpus within four word ranges: 5, 25, 100, and 500. That is to say, if a source and a target appear within five words of one another, this is included in the five-word proximity feature. This generates four features per relation.

2.3.3 Subjectivity keyword proximity counts

We augment the proximity counts feature with a third requirement: that the source and target appear within one of the ranges with a "subjectivity keyword." The keywords are taken from University of Pittsburgh subjectivity lexicon; the utility of this lexicon is supported in recent work (Somasundaran and Wiebe, 2009).

For performance reasons, we did not use all of the entries in the subjectivity lexicon. Instead, we used a TFIDF-based measure to rank the keywords by their prevalence in the ComputerWorld corpus where the term frequency is defined over the entire corpus. Then we selected 200 keywords with the highest score.

For each keyword, we use the same proximity ranges (5, 25, 100, and 500) in queries to Indri where we obtain counts of each keyword-source-target triple for each range. There are threfore 800 subjectivity keyword features.

Positive class	Negative class	System	Prec / Rec / F	Accuracy
Y	N	Random baseline	0.60 / 0.53 / 0.56	0.52
Y	N	Maj.-class (Y) baseline	0.59 / *1.00* / 0.74	0.59
Y	N	Linear kernel	0.70 / 0.73 / 0.72	0.66
Y	N	RBF kernel	**0.72 / 0.76 / 0.75**	**0.69**
Y	N/X	Random baseline	0.44 / 0.50 / 0.47	0.50
Y	N/X	RBF kernel	0.65 / 0.55 / 0.59	0.67

Table 1: Results with all features against majority class and random baselines. All values are mean averages under 10-fold cross validation.

2.3.4 Word context (unigram) features

For each relation, we take term frequency counts of the paragraph to which the relation belongs. We multiply them by the IDF of the term across the ComputerWorld corpus. This yields 6644 features over all paragraphs.

2.4 Machine learning

On these feature vectors, we trained SVM models using Joachims' (1999) svmlight tool. We use a radial basis function kernel with an error cost parameter of 100 and a γ of 0.25. We also use a linear kernel with an error cost parameter of 100 because it is straightforwardly possible with a linear kernel to extract the top features from the model generated by svmlight.

3 Experiments

We conducted most of our experiments with only the Y and N classes, discarding all X; this restricted most of our results to those assuming correct named entity recognition. Y was the positive class for training the svmlight models, and N was the negative class. We also performed experiments with N and X together being the negative class; this represents the condition that we are seeking "expresses-an-opinion-about" even with a higher named-entity error rate.

We use two baselines. One is a random baseline with uniform probability for the positive and negative classes. The other is a majority-class assigner (Y is the majority class).

The best system for the Y vs. N experiment was subjected to feature ablation. We first systematically removed each of the four feature types individually. The feature type whose removal had the largest effect on performance was removed permanently, and the rest of the features were tested without it. This was done once more, at which point only one feature type was present in the models tested.

3.1 Evaluation

All evaluation was performed under 10-fold cross validation, and we report the mean average of all performance metrics (precision, recall, harmonic mean F-measure, and accuracy) across folds.

We define these measures in the standard information retrieval form. If tp represents true positives, tn true negatives, fp false positives, and fn false negatives, then precision is $tp/(tp+fp)$, recall $tp/(tp+fn)$, F-measure (harmonic mean) is $2(prec * rec)/(prec + rec)$, and accuracy is $(tp + tn)/(tp + fp + fn + tn)$.

4 Results and discussion

The results of the experiments with all features are listed in table 1.

4.1 "Perfect" named entity recognition

We achieve best results in the Y versus N case using the radial basis function kernel. We find improvement in F-measure and accuracy at 19% and 17% respectively. Simply assigning the majority class to all test examples yields a very high recall, by definition, but poor precision and accuracy; hence its relatively high F-measure does not reflect high applicability to further processing, as the false positives would amplify errors in our social science application.

The linear kernel has results that are below the RBF kernel for all measures, but are relatively close to the RBF results.

Subjectivity	Proximity	Frequency	Unigram	Prec / Rec / F	Accuracy
✓	✓	✓	✓	0.72 / 0.76 / 0.75	0.69
	✓	✓	✓	0.67 / 0.89 / 0.76	0.67
✓		✓	✓	0.71 / 0.77 / 0.73	0.68
✓	✓		✓	0.70 / 0.78 / 0.74	0.67
✓	✓	✓		0.69 / 0.77 / 0.73	0.67
		✓	✓	0.63 / 0.91 / 0.75	0.64
	✓		✓	0.66 / 0.89 / 0.76	0.67
	✓	✓		0.65 / 0.90 / 0.76	0.66
		✓		0.61 / 0.92 / 0.73	0.60
			✓	0.61 / 0.94 / 0.74	0.60

Table 2: Feature ablation results for RBF kernel on `Y` vs. `N` case. The first line is the RBF result with all features from table 1.

4.2 Introducing erroneous named entities

The case of `Y` versus `N` and `X` together unsurprisingly performed worse than the case where named entity errors were eliminated. However, relative to its own random baseline, it performed well, with a 12% and 17% improvement in F-measure and accuracy using the RBF kernel. This suggests that the errors do not introduce enough noise into the system to produce a large decline in performance.

As `X` instances are about 26% of the total and we see a considerable drop in recall, we can say that some of the `X` instances are likely to be similar to valid `Y` ones; indeed, examination of the named entity recognizer's errors suggests that some incorrect organizations (e.g. product names) occur in contexts where valid organizations occur. However, precision and accuracy have not fallen nearly as far, so that the quality of the output for further processing is not hurt in proportion to the introduction of `X` class noise.

4.3 Feature ablation

Table 2 contains the result of our feature ablation experiments. Overall, the removal of features causes the SVM models to behave increasingly like a majority class assigner. As we mentioned earlier, higher recall at the expense of precision and accuracy is not an optimal outcome for us even if the F-measure is preserved. In our results, the F-measure values are remarkably stable.

In the first round of feature removal, the subjectivity keyword features have the biggest effect with the largest drop in precision and the largest increase in recall; high-TFIDF words from a general-purpose subjectivity lexicon allow the model to assign more items to the negative class.

The next round of feature removal shows that the proximity features have the next largest amount of influence on the classifier, as precision drops by 4%. The proximity features are very similar to the subjectivity features in that they too involve queries over windows of limited word sizes; the subjectivity keyword features only differ in that a subjectivity keyword must be within the window as well. That the proximity features are not more important than the subjectivity features, implies that the subjectivity keywords matter to the classifier, even though they are not specific to the IT domain. However, the proximity of sources and targets also matters, even in the absence of the subjectivity keywords.

Finally, we are left with the frequency features and the unigram context features. Either set of features supports a level of performance greater than the random baseline in table 1. However, the unigram features allow for slightly better recall than the frequency features without loss of precision, but this may not be very surprising, as there are *many* more unigram features than frequency features. More importantly, however, either of these feature types is sufficient to prevent the classifier from assigning the majority class all of the time, although they come close.

Feature type	Range	Keyword
Subjectivity	500	agreement
Subjectivity	500	critical
Subjectivity	500	want
Subjectivity	100	will
Subjectivity	100	able
Subjectivity	500	worth
Subjectivity	500	benefit
Subjectivity	100	trying
Subjectivity	500	large
Subjectivity	500	competitive

Table 3: The 10 most positive features via a linear kernel in descending order.

Feature type	Range	Keyword
Subjectivity	500	low
Subjectivity	500	ensure
Subjectivity	25	want
Subjectivity	100	vice
Subjectivity	500	slow
Subjectivity	100	large
Subjectivity	500	ready
Subjectivity	100	actually
Subjectivity	100	ready
Subjectivity	100	against

Table 4: The 10 most negative features via a linear kernel in descending order.

4.4 Most discriminative features

The models generated by svmlight under a linear kernel allow for the extraction of feature weights by a script written by svmlight's creator. We divided the instances into a single 70%/30% train/test split and trained a classifier with a linear kernel and an error cost parameter of 100, with results similar to those reported under 10-fold cross-validation in table 1. We used all features.

Then we were able to extract the 10 most positive (table 3) and 10 most negative (table 4) features from the model.

Interestingly, all of these are subjectivity keyword features, even the negatively weighted features. The top positive features are often evocative of business language, such as "agreement", "critical", and "competitive". Most of them emerge from queries at the 500-word range, suggesting that their presence in the document itself is evidence that a source is expressing an opinion about a target. That most of them are subjectivity features is reflected in the feature ablation results in the previous section.

It is less clear why "ensure" and "against" should be evidence that a source-target pair is *not* an instance of "expresses-an-opinion-about". On the other hand, words like "ready" (which appears twice) and "actually" can conceivably reflect situations in the IT domain that are not matters of opinion. In either case, this demonstrates one of the advantages of our technique, as these are features that actively assist in classifying some relation instances as not expressing sentiment. For example, contrary to what we would expect, "want" in a 25-word window with a source and a target is actually evidence against an "expresses-an-opinion-about" relation in text about IT innovations (ComputerWorld, July 2007):

> But **Klein**, who is director of information services and technology, didn't want IT to become the *blog* police.

In this example, Klein is expressing a desire, but not about the innovation (blogs) in question.

5 Conclusions and future work

5.1 Summary

We constructed and evaluated a system that detects at paragraph level whether entities relevant to the IT domain have expressed an opinion about a list of IT innovations of interest to a larger social science research program. To that end, we used a combination of co-occurrence statistics gleaned from a document indexing tool and TFIDF values from the local term context. Under these novel conditions, we successfully exceeded simple baselines by large margins.

Despite only moderate annotator agreement, we were able to produce results coherent enough to successfully train classifiers and conduct experiments.

Our feature ablation study suggests that all of the feature types played a role in improving the performance of the system over the random and majority-class baselines. However, the subjectivity keyword features from an existing lexicon played the largest role, followed by the proximity and unigram features. Subjectivity keyword features dominated the ranks of feature weights under a linear kernel, and the features most predictive of membership in "expresses-an-opinion-about" are words with semantic significance in the context of the IT business press.

5.2 Application to other domains

We used somewhat naïve statistics in a simple machine learning system in order to implement a form of opinion mining for a particular domain. The most direct linguistic guidance we provided our system were the query ranges and the subjectivity lexicon. The generality of this approach yields the advantage that it can be applied to other domains where there are ways of expressing sentiment unique to those domains outside of newswire text and product reviews.

5.3 Improving the features

Our use of an existing sentiment lexicon opens the door in future work for the use of techniques to bootstrap a larger sentiment lexicon that emphasizes domain-specific language in the expression of opinion, including opinionated acts. In fact, our results suggest that terminology in the existing lexicon that is most prominently weighted in our classifier also tends to be domain-relevant. In a further iteration, we might also improve performance by using terms outside the lexicon that tend to co-occur with terms from the lexicon.

5.4 Data generation

Our annotation exercise was a very simple one involving a short reading exercise and the selection of one of three choices per relation instance. This type of exercise is ideally suited to the "crowdsourcing" technique of paying many individuals small amounts of money to perform these simple annotations over the Internet. Previous research (Snow et al., 2008) suggests that we can generate very large datasets very quickly in this way; this is a requirement for expanding to other domains.

5.5 Scalability

In order to classify on the order of 1000 instances, it took nearly a million queries to the Indri index, which took a little over an hour to do in parallel on 25 processors by calling the Indri query engine afresh at each query. While each query is necessary to generate each feature value, there are a number of optimizations we could implement to accelerate the process. Various types of dynamic programming and caching could be used to handle related queries. One way of scaling up to larger datasets would be to use the MapReduce and cloud computing paradigms on which text processing tools have already been implemented (Moreira et al., 2007).

The application for this research is a social science exercise in exploring trends in IT adoption by analysing the IT business press. In the end, the perfect discovery of all instances of "expresses-an-opinion-about" is not as important as finding enough reliable data over a large number of documents. This work brings us several steps closer in finding the right combination of features in order to acquire trend-representative data.

Acknowledgements

This paper is based upon work supported by the National Science Foundation under Grant IIS-0729459.

References

Bikel, Daniel M., Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Mach. Learn.*, 34(1-3).

Choi, Yejin, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Girju, Roxana, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: classification of semantic relations between nominals. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18, Morristown, NJ, USA. Association for Computational Linguistics.

Joachims, T. 1999. Making large-scale SVM learning practical. In Schölkopf, B., C. Burges, and

A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

Kim, Soo-Min and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Metzler, Donald and W. Bruce Croft. 2004. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735 – 750.

Mooney, Raymond J. and Razvan Bunescu. 2005. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10.

Moreira, José E., Maged M. Michael, Dilma Da Silva, Doron Shiloach, Parijat Dube, and Li Zhang. 2007. Scalability of the nutch search engine. In Smith, Burton J., editor, *ICS*, pages 3–12. ACM.

Rogers, Everett M. 2003. *Diffusion of Innovations, 5th Edition*. Free Press.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP 2008*, Morristown, NJ, USA.

Somasundaran, Swapna and Janyce Wiebe. 2009. Recognizing stances in online debates. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. Association for Computational Linguistics.

Wilson, Theresa and Janyce Wiebe. 2005. Annotating attributions and private states. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*, pages 53–60.

Wilson, Theresa, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005a. OpinionFinder: A system for subjectivity analysis. In *HLT/EMNLP*.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*.