A Cross-corpus Study of Unsupervised Subjectivity Identification based on Calibrated EM

Dong Wang Yang Liu
The University of Texas at Dallas
{dongwang,yangl}@hlt.utdallas.edu

Abstract

In this study we investigate using an unsupervised generative learning method for subjectivity detection in text across different domains. We create an initial training set using simple lexicon information, and then evaluate a calibrated EM (expectation-maximization) method to learn from unannotated data. We evaluate this unsupervised learning approach on three different domains: movie data, news resource, and meeting dialogues. We also perform a thorough analysis to examine impacting factors on unsupervised learning, such as the size and self-labeling accuracy of the initial training set. Our experiments and analysis show inherent differences across domains and performance gain from calibration in EM.

1 Introduction

Subjectivity identification is to identify whether an expression contains opinion or sentiment. Automatic subjectivity identification can benefit many natural language processing (NLP) tasks. For example, information retrieval systems can provide affective or informative articles separately (Pang and Lee, 2008). Summarization systems may want to summarize factual and opinionated content differently (Murray and Carenini, 2008). In this paper, we perform subjectivity detection at sentence level, which is more appropriate for some subsequent processing such as opinion summarization.

Previous work has shown that when enough labeled data is available, supervised classification methods can achieve high accuracy for subjectivity detection in some domains. However, it is often expensive to create such training data. On the other hand, a lot of unannotated data is readily available in various domains. Therefore an interesting and important problem is to develop semi-supervised or unsupervised learning methods that can learn from an unannotated corpus. In this study, we use an unsupervised learning approach where we first use a knowledge-based method to create an initial training set, and then apply a calibrated EM method to learn from an unannotated corpus. Our experiments show significant differences among the three domains: movie, news article, and meeting dialog. This can be explained by the inherent difference of the data, especially the task difficulty and classifier's performance for a domain. We demonstrate that for some domains (e.g., movie data) the unsupervised learning methods can rival the supervised approach.

2 Related Work

In the early age, knowledge-based methods were widely used for subjectivity detection. They used a lexicon or patterns and rules to predict whether a target is subjective or not. These methods tended to yield a high precision and low recall, or low precision and high recall (Kim and Hovy, 2005). Recently, machine learning approaches have been adopted more often (Ng et al., 2006). There are limitations in both methods. In knowledge-based approaches, a predefined subjectivity lexicon may not adapt well to different domains. While in machine learning approach, human labeling efforts are required to create a large training set.

To overcome the above drawbacks, unsupervised or semi-supervised methods have been explored in sentiment analysis. For polarity classification, some previous work used spectral techniques (Dasgupta and Ng, 2009) or co-training (Li et al., 2010) to mine the reviews in a semi-supervised manner. For subjectivity identification, Wiebe and Riloff (Wiebe and Riloff, 2005) applied a rule-based method to create a training set first and then used it to train a naive Bayes classifier. Melville et al. (Melville et al., 2009) used a pooling multinomial method to combine lexicon derived probability and statistical probability.

Our work is similar to the study in (Wiebe and Riloff, 2005) in that we both use a rule-based method to create an initial training set and learn from

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 161–167,
24 June, 2011, Portland, Oregon, USA

unannotated corpus. However, there are two key differences. First, unlike the self-training method they used, we use a calibrated EM iterative learning approach. Second, we compare the results on three different corpora in order to evaluate the domain/genre effect of the unsupervised method. Our cross-corpus study shows how the unsupervised learning approach performs in different domains and helps us understand what are the factors impacting the learning methods.

3 Data

We use three data sets from different domains: movie, news resource, and meeting conversations. The first two are from written text domain and have been widely used in many previous studies for sentiment analysis (Pang and Lee, 2004; Raaijmakers and Kraaij, 2008). The third one is from speech transcripts. It has been used in a few recent studies (Raaijmakers et al., 2008; Murray and Carenini, 2009), but not as much as those text data. The following provides more details of the data.

- The first corpus is movie data (Pang and Lee, 2004). It contains 5,000 subjective sentences collected from movie reviews and 5,000 objective sentences collected from movie plot summaries. The sentences in each collection are randomly ordered.
- The second one is extracted from MPQA corpus (version 2.0) (Wilson and Wiebe, 2003), which is collected from news articles. This data has been annotated with subjective information at phrase level. We adopted the same rules as in (Riloff and Wiebe, 2003) to create the sentence level label: if a sentence has at least one private state of strength medium or higher, then the sentence is labeled SUBJECTIVE, otherwise it is labeled OBJECTIVE. We randomly extracted 5,000 subjective and 5,000 objective sentences from this corpus to make it comparable with the movie data.
- The third data set is from AMI meeting corpus. It has been annotated using the scheme described in (Wilson, 2008). There are 3 main categories of annotations regarding sentiments: subjective utterances, subjective questions, and objective polar utterances. We consider the union of subjective utterance and subjective question as subjective and the rest as objective. The subjectivity classification task is done at the dialog act (DA) levels. We label each DA using the label of the utterance that has overlap with it. We create a balanced data set using this corpus, containing 9,892 DAs in total. This number is slightly less than those for movie and MPQA data because of the available data size in this corpus. The data is also randomly ordered without considering the role of the speaker and which meeting it belongs to.

Table 1 summarizes statistics for the three data sets. We can see that sentences in meeting dialogs (AMI data) are generally shorter than the other domains, and that sentences in news domain (MPQA) are longer, and also have a larger variance. In addition, the inter-annotator agreement on AMI data is quite low, which shows it is even difficult for human to determine whether an utterance contains sentiment in meeting conversations.

		Movie	MPQA	AMI
sent length	**min**	3	1	3
	max	100	246	67
	mean	20.37	22.38	8.78
	variance	75.26	147.18	34.26
vocabulary size		15,847	13,414	3,337
Inter-annotator agreement		N/A	0.77	0.56

Table 1: Statistics for the three data sets: movie, MPQA, and AMI data. The inter-annotator agreement on movie data is not available because it is not annotated by human.

4 Unsupervised Subjectivity Detection

In this section, we describe our unsupervised learning process that uses a knowledge-based method to create an initial training set, and then uses a calibrated EM approach to incorporate unannotated data into the learning process. We use a naive Bayes classifier as the base supervised classifier with a bag-of-words model.

4.1 Create Initial Training Set

A lexicon-based method is used to create an initial training set, since it can often achieve high precision rate (though low recall) for subjectivity detection. We use a subjectivity lexicon (Wilson et al., 2005) to calculate the subjectivity score for each sentence.

This lexicon contains 8,221 entries that are categorized into strong and weak subjective clues.

For each word w, we assign a subjectivity score $sub(w)$: 1 to strong subjective clues, 0.5 to weak clues, and 0 for any other word. Then the subjectivity score of a sentence is the sum of the values of all the words in the sentence, normalized by the sentence length. We noticed that for sentences labeled as SUBJECTIVE in the three corpora, the subjective clues appear more frequently in movie data than the other two corpora. Thus we perform different normalization for the three data sets to obtain the subjectivity score for each sentence, $sub(s)$: Equation 1 for the movie data, and Equation 2 for MPQA and AMI data.

$$sub(s) = \sum_{w \in s} sub(w)/sent\_length \quad (1)$$

$$sub(s) = \sum_{w \in s} sub(w)/log(sent\_length) \quad (2)$$

We label the top m sentences with the highest subjective scores as SUBJECTIVE, and label m sentences with the lowest scores as OBJECTIVE. These $2m$ sentences form the initial training set for the iterative learning methods.

4.2 Calibrated EM Naive Bayes

Expectation-Maximization (EM) naive Bayes method is a semi-supervised algorithm proposed in (Nigam et al., 2000) for learning from both labeled and unlabeled data. In the implementation of EM, we iterate the E-step and M-step until model parameters converge or a predefined iteration number is reached. In E-step, we use naive Bayes classifier to estimate the posterior probabilities of each sentence s_i belonging to each class c_j (SUBJECTIVE and OBJECTIVE), $P(c_j|s_i)$:

$$P(c_j|s_i) = \frac{P(c_j)\prod_{k=1}^{|s_i|} P(w_k|c_j)}{\sum_{c_l \in C} P(c_l)\prod_{k=1}^{|s_i|} P(w_k|c_l)} \quad (3)$$

The M-step uses the probabilistic results from the E-step to recalculate the parameters in the naive Bayes classifier, the probability of word w_t in class c_j and the prior probability of class c_j:

$$P(w_t|c_j) = \frac{0.1 + \sum_{s_i \in S} N(w_t, s_i) P(c_j|s_i)}{0.1 \times |V| + \sum_{k=1}^{|V|} \sum_{s_i \in S} N(w_k, s_i) P(c_j|s_i)} \quad (4)$$

$$P(c_j) = \frac{0.1 + \sum_{s_i \in S} P(c_j|s_i)}{0.1 \times |C| + |S|} \quad (5)$$

S is the set of sentences. $N(w_t, s_i)$ is the count of word w_t in a sentence s_i. We use additive smoothing with $\alpha = 0.1$ for probability parameter estimation. $|C|$ is the number of classes, which is 2 in our case, and $|V|$ is the vocabulary size, obtained from the entire data set.

In the first iteration, we assign $P(c_j|s_i)$ using the pseudo training data generated based on lexicon information. If a sentence is labeled SUBJECTIVE, then $P(sub|s_i)$ is 1 and $P(obj|s_i)$ is 0; for the sentences with OBJECTIVE labels, $P(sub|s_i)$ is 0 and $P(obj|s_i)$ is 1.

In our work, we use a variant of standard EM: calibrated EM, introduced by (Tsuruoka and Tsujii, 2003). The basic idea of this approach is to shift the probability values of unlabeled data to the extent such that the class distribution of unlabeled data is identical to the distribution in labeled data (balanced class in our case). In our approach, before model training ("M-step") in each iteration, we adjust the posterior probability of each sentence in the following steps:

- Transform the posterior probabilities through the inverse function of the sigmoid function. The outputs are real values.
- Sort them and use the median of all the values as the border value. This is because our data is balanced.
- Subtract this border value from the transformed values.
- Transform the new values back into probability values using a sigmoid function.

Note that there is a caveat here. We are assuming we know the class distribution, based on labeled training data or human knowledge. This is often a reasonable assumption. In addition, we are assuming that this class distribution is the same for the unlabeled data. If this is not true, then the distribution adjustment performed in calibrated EM may hurt system performance.

5 Empirical Evaluation

In this section, we evaluate our unsupervised learning method and analyze various impacting factors.

In preprocessing, we removed the punctuation and numbers from the data and performed word stemming. To measure performance, we use classification accuracy.

5.1 Unsupervised Learning Results

In experiments of unsupervised learning, we perform 5-fold cross validation. We divide the corpus into 5 parts with equal size (each with balanced class distribution). In each run we reserve one part as the test set. From the remaining data, we use the lexicon-based method to create the initial training data, containing 1,000 SUBJECTIVE and 1,000 OBJECTIVE sentences. The rest is used as unlabeled data to perform iterative learning. The final model is then applied to the reserved test set. Figure 1 shows the learning curves of calibrated EM on movie, MPQA and AMI data respectively.

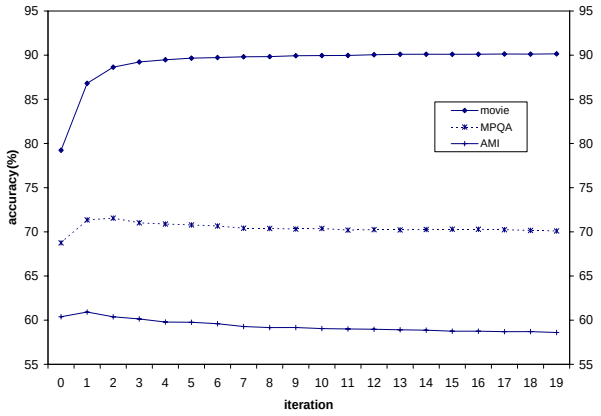

Figure 1: Calibrated EM results using unsupervised setting (2,000 self-labeled initial samples) on movie, MPQA, and AMI data.

On movie data, calibrated EM improves the performance significantly ($p<0.005$), compared to that based on the initial training set (iteration 0). It takes only a few iterations for the EM method to converge and at the end of the iteration, it achieves 90.15% accuracy, which rivals the fully supervised learning performance (91.31% when using all the 8,000 labeled sentences for training). On MPQA data, this method yields some improvement ($p<0.1$) compared to the initial point. But there is a peak accuracy in the first couple of iterations, and then performance starts dropping thereafter. On AMI data, the performance degrades after the first iteration.

5.2 Analysis and Discussion

5.2.1 Effect of initial set

For unsupervised learning, our first question is how the accuracy and size of the initial training set affect performance. We calculate the self-labeling accuracy for the initial set using the lexicon based method. Table 2 shows the labeling accuracy when using different initial size, measured for SUBJECTIVE and OBJECTIVE class separately. In addition, we present the classification performance on the test set when using the naive Bayes classifier trained from the initial set. Each size in the table represents the total number of sentences in the initial set.

Table 2 shows that when the size is 2,000 (as we used in previous experiments), the accuracy for both classes on MPQA are even better than on movies, even though we have seen that iterative learning methods perform much better on movies, suggesting that the initial data set accuracy is not the reason for the worse performance on MPQA than movies. It also shows that on movie data, as the initial size increases, the accuracy of the pseudo training set decreases, which is as expected (the top ranked self-labeled samples are more confident and accurate). However, this is not the case on MPQA and AMI data. There is no obvious drop of accuracy, rather in many cases accuracy even increases when the initial size increases. It shows that on these two corpora, our lexicon-based method does not perform very well because the most highly ranked sentences according to the subjective lexicon are not those most subjective sentences.

size		**100**	**200**	**1000**	**2000**	**3000**
movie	**sub**	95.20	92.20	82.48	79.24	77.13
	obj	82.20	82.00	80.88	79.04	77.31
	Acc_Test	59.93	71.63	77.62	79.24	79.64
MPQA	**sub**	83.20	85.60	85.76	85.18	82.53
	obj	87.60	86.60	87.64	87.46	85.92
	Acc_Test	60.45	63.83	66.98	68.75	70.05
AMI	**sub**	49.60	53.40	65.96	66.98	67.05
	obj	71.60	71.00	68.56	69.04	69.89
	Acc_Test	50.51	53.81	60.53	60.39	60.46

Table 2: Initial pseudo training accuracy for SUBJECTIVE (sub) and OBJECTIVE (obj) class, and performance on the test using this initial training set (Acc_Test). Results (all in %) are shown for different initial data size.

From the results on the test set, we find that when

the size is smaller, such as containing 100 or 200 samples, the accuracy on test set is lower than using a bigger initial set. This is mainly because there is not sufficient data for model training. For AMI data, this is also due to the low accuracy in the training set. When the initial size is large enough, the improvement from a larger training set is not as substantial, for example, using 1,000, 2,000, or 3,000 sentences. On AMI data, there is almost no difference among the three sets. There is a tradeoff between the two factors, self-labeling accuracy and the data size. Often an improvement in one aspect causes degradation of the other. A reasonable starting point needs to be chosen considering both factors. Overall, it shows that the performance on test set can benefit more from using a larger initial training set, though it may be noisy.

In order to further investigate the impact of self-labeled initial data set, we perform standard semi-supervised learning using reference labels in the initial data set. The learning curve of this semi-supervised setting is shown in Figure 2.

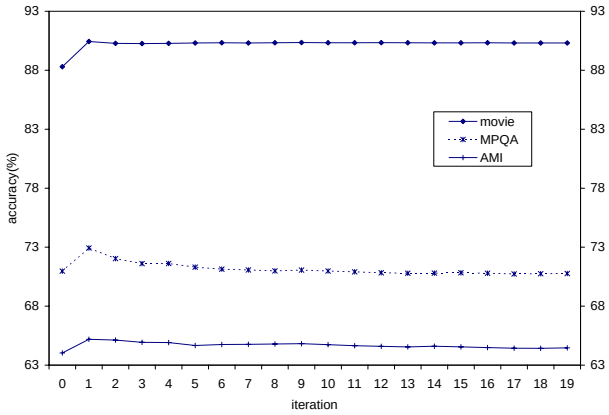

Figure 2: Calibrated EM results using semi-supervised learning (2,000 labeled seed) on movie, MPQA, and AMI data.

On movie data, calibrated EM yields better performance over that based on the initial training data (iteration 0). We can see that calibrated EM converges very fast and achieves very high performance in the first iteration. On MPQA and AMI data, calibrated EM increases the accuracy at the first iteration but then degrades thereafter. This shows that incorporating unlabeled data in training is helpful, however, more EM iterations do not yield further gain.

We noticed that on AMI data, even when the initial set has 100% accuracy (i.e., semi-supervised setting), it still fails to yield any performance gain on AMI data. It shows that the low accuracy of initial training set does not explain the poor performance of unsupervised learning method. Therefore, we conducted another set of experiments which use the same semi-supervised setting but start from different initial training sizes. We observed that on MPQA and AMI data, calibrated EM is able to increase the accuracy only when the initial training set is small (less than 100 instances) and the performance at the start point is poor. We believe this is related to the data property and the assumptions used in EM. Similar patterns have been found in some previous studies (Chapelle et al., 2006). They attribute this to the incorrect model assumption, i.e., when the modeling assumptions for a particular classifier do not match the characteristics of the distribution of the data, unlabeled data may degrade the performance of classifiers.

5.2.2 Effect of calibration

Figure 3 compares calibrated EM with standard EM using unsupervised learning on the three domains. We can see that calibrated EM outperforms standard EM, with a larger improvement on MPQA and AMI data. When using standard EM, we find that there is a larger difference between the number of instances in the two classes based on the model's prediction on MPQA and AMI data than movie data. For example, in one run using EM, in the first iteration the ratio of the two classes is 2.21, 1.88, and 1.23 for MPQA, AMI, and movie data respectively. Calibrated EM is more effective on the two domains because it adjusts the posterior probability of each sample according to the class distribution in the data, making it more accurate in training the model in the next iteration.

5.2.3 Error analysis

There are two points worth discussing based on our error analysis.

A. Domain difference.

Much of the difference we have observed can be attributed to the genre difference. In movie reviews, often a person expresses his/her favor (or not) of the movie explicitly, making the task relatively easy for automatic subjectivity classification. MPQA data is collected from news resource, where subjectivity mostly means an attitude or a judgment. Take

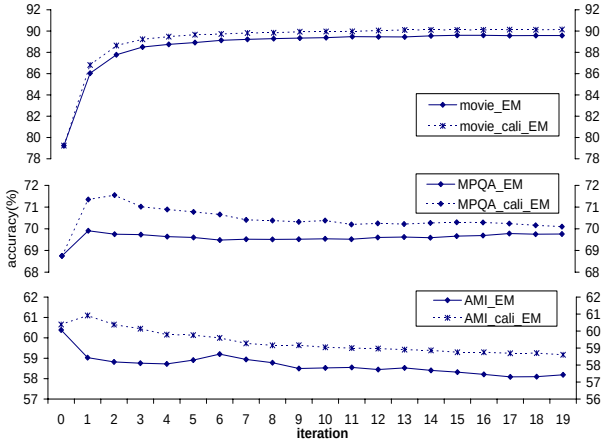

Figure 3: Comparison of standard EM and calibrated EM.

the following sentence as an example: "The United States is prepared to fight terrorism alone". It is labeled as SUBJECTIVE because it expresses a determination. However, it may also be interpreted as an objective statement.

The AMI corpus consists of meeting conversations. The free-style dialogues are very different from the style in review and news articles. There are many incomplete sentences and disfluencies. More importantly, the meaning of a sentence is often context dependent. In the examples shown below, the two sentences look very similar, however, the first sentence is labeled as "OBJECTIVE", and the second one as "SUBJECTIVE". This is because of the different context and speaker information – the second sentence expresses agreement, but the first example is just a sequence of discourse marker words.

- Alright yeah okay
- Yeah okay, true, true.

We notice that many of the classification errors in AMI occur in very short sentences, like in the example shown above. These short sentences are very ambiguous for subjectivity classification.

B. Limitation of the bag-of-word model.

Our analysis also showed that some sentences are difficult to classify if simply using surface words. In the following, we show some examples of system errors.

False negatives: subjective sentences recognized as objective

- Johnson has, in his first film, set himself a task he is not nearly up to. (movie data)
- The news from Israel is almost earth-shattering. (MPQA)
- We can stick with what we already get. (AMI)

False positives: objective sentences recognized as subjective

- Cathy (Julianne Moore) is the **perfect** 50s housewife, living the **perfect** 50s life: **healthy** kids, **successful** husband, social **prominence**. (movie data)
- The committee Wednesday opened a formal debate on human rights questions, including alternative approaches for **improving** the **effective** enjoyment of human rights and **fundamental freedoms**. (MPQA)
- um uh you know apple been really **successful** with this surgical white kind of business or this **sleek** kind of (AMI)

In the first three examples, there are no explicit subjective clues, resulting in false negative errors. The subjective word "earth-shattering" is not included in subjective lexicon and rarely used in the corpus. The last three examples contain several subjective words, and are therefore labeled as subjective. These are the problems with the current word based approaches.

6 Conclusion and Future Work

This paper investigates an unsupervised learning procedure for subjectivity identification at sentence level. We use a lexicon-based method to create initial training data and then apply a calibrated EM to utilize unlabeled corpus. We evaluate this method across three different data sets and observe significant difference. It yields good performance on movie data but does not achieve much performance gain on MPQA corpus, while on AMI corpus it fails to yield improvement. Our analysis showed that performance of the base classifier has a substantial impact on iterative learning methods. In addition, we found that calibrated EM outperforms the standard EM method when the class distribution based on classifier's hypotheses does not match the real one.

Our iterative learning approach uses a naive Bayes classifier that may not have accurate posterior probabilities. Therefore in our future work, we will evaluate using other base models. Our cross-corpus analysis shows poor performance of subjectivity detection in AMI data. We plan to explore more information from multiparty dialogs to help improve performance for that domain.

7 Acknowledgment

The authors thank Theresa Wilson for sharing annotation for the AMI corpus and helping with data processing for that data. Part of this work is supported by an NSF award CNS-1059226.

References

O. Chapelle, B. Schölkopf, and A. Zien, editors. 2006. Semi-supervised learning. MIT Press.

Sajib Dasgupta and Vincent Ng. 2009. *Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification*. In *Proceedings of ACL-IJCNLP*, pages 701–709.

Soo-Min Kim and Eduard Hovy. 2005. *Automatic detection of opinion bearing words and sentences*. In *Proceedings of ACL*.

Shoushan Li, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. 2010. *Employing personal/impersonal views in supervised and semi-supervised sentiment classification*. In *Proceedings of ACL*, pages 414–423.

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. 2009. *Sentiment analysis of blogs by combining lexical knowledge with text classification*. In *Proceedings of ACM SIGKDD*, pages 1275–1284.

Gabriel Murray and Giuseppe Carenini. 2008. *Summarizing spoken and written conversations*. In *Proceedings of EMNLP*, pages 773–782.

Gabriel Murray and Giuseppe Carenini. 2009. *Detecting subjectivity in multiparty speech*. In *Proceedings of Interspeech*.

Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. *Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews*. In *Proceedings of COLING/ACL*, pages 611–618.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. *Text classification from labeled and unlabeled documents using EM*. *Machine Learning*, 39:103–134.

Bo Pang and Lilian Lee. 2004. *A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts*. In *Proceedings of ACL*, pages 271–278.

Bo Pang and Lillian Lee. 2008. *Using very simple statistics for review search: An exploration*. In *Proceedings of COLING*, pages 73–76.

Stephan Raaijmakers and Wessel Kraaij. 2008. *A Shallow approach to subjectivity classification*. In *Proceedings of ICWSM*.

Stephan Raaijmakers, Khiet Truong, and Theresa Wilson. 2008. *Multimodal subjectivity analysis of multiparty conversation*. In *Proceedings of EMNLP*, pages 466–474.

Ellen Riloff and Janyce Wiebe. 2003. *Learning extraction patterns for subjective expressions*. In *Proceedings of EMNLP*, pages 105–112.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2003. *Training a naive bayes classifier via the EM algorithm with a class distribution constraint*. In *Proceedings of NAACL*, pages 127–134.

Janyce Wiebe and Ellen Riloff. 2005. *Creating subjective and objective sentence classifiers from unannotated texts*. In *Proceedings of CICLing*, pages 486–497.

Theresa Wilson and Janyce Wiebe. 2003. *Annotating opinions in the world press*. In *Proceedings of SIGdial*, pages 13–22.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. *Recognizing contextual polarity in phrase-level sentiment analysis*. In *Proceedings of HLT-EMNLP*, pages 347–354.

Theresa Wilson. 2008. *Annotating subjective content in meetings*. In *Proceedings of LREC*.