Natural Language Engineering **18** (3): 375–397. © Cambridge University Press 2011
doi:10.1017/S1351324911000234

A cross-corpus study of subjectivity identification using unsupervised learning†

DONG WANG and YANG LIU

Department of Computer Science, The University of Texas at Dallas,
800 West Campbell Road, Richardson, Texas
e-mail: {dongwang, yangl}@hlt.utdallas.edu

(*Received 25 October 2010; revised 27 April 2011; accepted 6 June 2011; first published online 16 August 2011*)

Abstract

In this study, we investigate using unsupervised generative learning methods for subjectivity detection across different domains. We create an initial training set using simple lexicon information and then evaluate two iterative learning methods with a base naive Bayes classifier to learn from unannotated data. The first method is self-training, which adds instances with high confidence into the training set in each iteration. The second is a calibrated EM (expectation-maximization) method where we calibrate the posterior probabilities from EM such that the class distribution is similar to that in the real data. We evaluate both approaches on three different domains: movie data, news resource, and meeting dialogues, and we found that in some cases the unsupervised learning methods can achieve performance close to the fully supervised setup. We perform a thorough analysis to examine factors, such as self-labeling accuracy of the initial training set in unsupervised learning, the accuracy of the added examples in self-training, and the size of the initial training set in different methods. Our experiments and analysis show inherent differences across domains and impacting factors explaining the model behaviors.

1 Introduction

The task of subjectivity identification is to identify whether an expression contains opinion, evaluation, or sentiments. Automatic subjectivity identification can benefit many natural language processing tasks. For example, information retrieval systems can provide affective or informative articles separately (Pang and Lee 2008). In question-answering, subjectivity identification can help answer more complex questions regarding the opinions expressed in a document (Yu and Hatzivassiloglou 2003). Summarization systems may want to summarize factual and opinionated content differently (Hu and Liu 2006; Murray and Carenini 2008; Nishikawa *et al.* 2010). For polarity classification, better performance can be achieved by removing the objective instances first (Ni *et al.* 2007).

There has been an increasing interest in subjectivity identification. Previous work has shown that when enough labeled data are available, supervised classification

† The authors thank Theresa Wilson for sharing annotation for the AMI corpus and helping with data processing for that data, and the three reviewers for their useful comments.

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

methods can achieve high accuracy for subjectivity detection in some domains. However, it is often expensive to create such training data. In addition, we may need a large amount of annotated data to have a good coverage for different domains. On the other hand, a lot of unannotated data are readily available in various domains. Therefore, an interesting and important problem is to develop semi-supervised or unsupervised learning methods that can learn from unannotated corpus. There have been some efforts in this direction for many natural language processing problems, including sentiment analysis. In this study, we use an unsupervised learning framework where we first use a knowledge-based method to create an initial training set and then apply iterative learning methods to use unlabeled data. We compare two iterative methods, self-training and calibrated EM. They both use a base naive Bayes (NB) classifier, but differ in their iterative learning process. We evaluate these methods using three different corpora: movie, news article, and meeting dialog, and we examine various factors that impact learning performance. Our experiments show significant differences among the three domains, which 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.

The rest of this paper is organized as follows. Section 2 briefly discusses related work. Section 3 describes the three data sets used in this study. We explain the unsupervised learning methods in Section 4 and present the experimental results and analysis in Section 5. Section 6 concludes the paper.

2 Related work

In previous work, subjectivity detection has been performed at different levels of granularity, including word level (Sebastiani, Esuli and Sebastiani 2006; Gyamfi *et al.* 2009), phrase level (Wilson, Wiebe and Hwa 2004), sentence level (Kim and Hovy 2005; Riloff *et al.* 2005), and document level (Yu and Hatzivassiloglou 2003; Wiebe *et al.* 2004; Pang and Lee 2008). In this paper, we perform subjectivity detection at sentence level. We expect that sentence-level decision is more appropriate for some subsequent processing, such as opinionated document retrieval in weblogs, opinion summarization.

In the early age, knowledge-based methods were widely used for subjectivity detection. They used a lexicon or patterns and a set of rules to predict whether a target is subjective or not. These methods tended to yield a high precision and low recall, or a low precision and high recall (Kim and Hovy 2005). Recently, machine learning approaches have been adopted more often (Ng, Dasgupta and Arifin 2006; Murray and Carenini 2009; Nakagawa, Inui and Kurohashi 2010). There are limitations in both knowledge-based and supervised learning methods. In knowledge-based approaches, a predefined subjectivity lexicon may not adapt well to different domains. Take the word 'lightweight' as an example. It may express positive meaning when describing a laptop, but is more neutral in a general lexicon. Machine learning approaches also suffer from this domain mismatch. Moreover,

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

human labeling efforts are required to create a large training set for machine learning approaches to work well.

To overcome the above drawbacks, unsupervised or semi-supervised methods have been explored in many classification problems to minimize human annotation and recently been used in sentiment analysis. For polarity classification, Dasgupta and Ng (2009) used spectral techniques to mine the unambiguous reviews first and then exploit them to classify the ambiguous reviews via a semi-supervised learning approach. Zhou, Chen and Wang (2010) performed active learning in a semi-supervised framework by identifying the instances that should be labeled as training data first and then labeling all unlabeled data using semi-supervised learning. Li *et al.* (2010) used co-training for product review data where each review can be thought of having two different views: personal and impersonal views. They reported good results compared to using rule-based classifiers or supervised approaches. Wiegand and Klakow (2010) used self-training and a rule-based method on polarity classification on review data and showed that their method can effectively boost the performance. For subjectivity identification, Wiebe and Riloff (2005) applied a rule-based method to an unlabeled corpus to create a training set and bootstrapped from unannotated data using self-training with a NB classifier. Melville, Gryc and Lawrence (2009) used a pooling multinomial method to combine lexicon-derived probability and statistical probability learned from a small amount of training data.

Our work is similar to Wiebe and Riloff (2005) and Wiegand and Klakow (2010) in the general framework in that we both use a simple rule-based method to create an initial training set and use self-training to learn from unannotated corpus. However, there are several key differences. We do not use context information as in Wiebe and Riloff (2005), because our task is sentence-level classification and we do not assume that sentences in our data are consecutive. We do not use a separate unannotated training set, but rather perform unsupervised learning using the test set itself. In addition, we compare two different iterative learning methods: self-training and calibrated EM. Furthermore, our study is conducted using three different corpora in order to evaluate the domain/genre effect of the unsupervised methods.

3 Data

We use three data sets from different domains: movie, news resource, and meeting conversations. The first two are from written text domains and have been widely used in many previous studies for subjectivity or polarity detection (Pang and Lee 2004; Kim and Hovy 2005; Raaijmakers and Kraaij 2008). The third one is from speech transcripts. It has been used in a few recent studies (Raaijmakers, Truong and Wilson 2008; Murray and Carenini 2009), but not as much as those text data. Spoken dialogs have inherent differences with written text in their styles. This cross-corpus study will show how the unsupervised learning approaches perform in different domains and help us understand what are the factors impacting the learning methods. The following provides more details of the data and shows an example of a subjective and an objective sentence for each data set:

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

- 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.

  Example:

  — It's hard to tell with all the crashing and banging where the salesmanship ends and the movie begins. (SUBJECTIVE)
  — The movie begins in the past where a young boy named sam attempts to save celebi from a hunter. (OBJECTIVE)

- The second one is extracted from MPQA corpus (version 2.0) (Wilson and Wiebe 2003), which is collected from news articles covering a variety of topics. This data have 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 to the movie data and avoid any impact due to the class distribution or data size.

  Example:

  — The world community should not tolerate crime of war. (SUBJECTIVE)
  — The European Commission announced it had pledged a financial package of grants and loans totaling 530 million euros (450 million dollars). (OBJECTIVE)

- The third data set is from AMI meeting corpus.[1] The meetings have been annotated using the scheme described in Wilson (2008). The annotation scheme was adapted from that used for MPQA, but included some conversation-related labels, such as questions. There are three main categories of annotations regarding sentiments: subjective utterances, subjective questions, and objective polar utterances. Our study aims to detect subjective content, so we consider the union of subjective utterance and subjective question as subjective, and the rest as objective. In the annotation, an utterance is a span of words that can be either within a dialogue act (DA) segment or cross two DAs. We label each DA using the label of the utterance that has overlap with it. The subjectivity classification task is done at these DA levels.[2] We create a balanced data set using this corpus, containing 9,892 DAs in total. We keep the minority class (OBJECTIVE) in the data and randomly sample the majority class to obtain the balanced data set. We did not consider speaker or meeting information during random sentence selection. The number of sentences in this corpus is slightly less than those for movie and MPQA data because of the available data.

[1] http://www.amiproject.org
[2] These DA segments are similar to sentences; therefore, we say that this study on subjectivity detection is at sentence level.

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

Table 1. *Statistics for the three data sets: movie, MPQA, and AMI data*

		Movie	MPQA	AMI
# of sents	subjective	5,000	5,000	4,946
	objective	5,000	5,000	4,946
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
interannotator agreement		N/A	0.77	0.56

Example:

— It does make sense from maybe the design point of view. (SUBJECTIVE)
— My task was this time to put up a questionnaire. (OBJECTIVE)

Table 1 summarizes statistics for these three data sets. For the movie data, since the data were collected from two sources: reviews and plots, it was not annotated like the other corpora; therefore, there is no interannotator agreement information for that data. We can see that in meeting dialogs (AMI data), sentences are generally shorter than the other domains, and sentences in news domain (MPQA) are longer, and sentences also have a larger variance. The vocabulary size is much bigger in movie and MPQA (especially the movie data considering that it has fewer tokens than MPQA) than that in AMI data. In addition, the interannotator 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. It is worth pointing out that since the original annotation is performed at frame level or utterance level for MPQA and AMI data, even though we use the same method as some previous work to obtain the sentence-level labels, these may not be perfect.

Note that the sentences in the original movie and MPQA data sets are randomly ordered, and thus, we cannot use information from other sentence context or discourse to determine the subjectivity of a sentence. For the AMI data, we adopt a similar setup as the other two corpora—the subjectivity detection task is sentence based, and the decision for each sentence is independent of other sentences.

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, similar to the work in Wiebe and Riloff (2005), and then use two iterative learning methods: self-training and calibrated EM, in order to incorporate unannotated data into the learning process. For the base supervised classifier in both learning methods, we choose to use a NB

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

classifier with a bag-of-words model, since this has been shown to perform well in previous work for sentiment analysis and other text categorization tasks (e.g., Pang, Lee and Vaithyanathan 2002; Yu and Hatzivassiloglou 2003). These unsupervised methods are iterative learning processes, different from general unsupervised learning methods, such as clustering.

4.1 Lexicon-based method to generate initial training set

To create an initial training set, we use a lexicon-based method, since they can often achieve high precision rate (though low recall) for subjectivity detection. We use a list of subjectivity clues from Wilson, Wiebe and Hoffmann (2005) to calculate the subjectivity score for each sentence. The lexicon is extracted from both written and spoken language, such as Wall Street Journal, British National Corpus, and speech corpus, and thus, it is considered as a general domain lexicon. It contains 8,221 entries that are categorized into strong and weak subjective clues, where strong ones are more likely to have a subjective meaning.

For each word w, we assign a subjective score $sub(w)$: 1 to strong subjective clues, 0.5 to weak clues, and 0 for any word that is not in the subjective lexicon. Then, the subjectivity score of a sentence is the sum of the values of all the words appearing in the sentence, normalized by the length of sentence. We noticed that for sentences labeled as SUBJECTIVE in the three corpora, the subjective clues appear relatively more frequently in movie data than in MPQA or AMI corpus. Thus, we perform different normalization for the three data sets to obtain the final subjective score for each sentence, $sub(s)$: Equation (1) for the movie data, and Equation (2) for MPQA and AMI data.[3]

$$sub(s) = \left(\sum_{w \in s}(sub(w))\right) / sent_length \tag{1}$$

$$sub(s) = \left(\sum_{w \in s} sub(w)\right) / log(sent_length) \tag{2}$$

After we calculate the subjectivity scores for all the sentences in the data set, we label the top m sentences with the highest subjective scores as SUBJECTIVE, and label m sentences with the lowest scores as OBJECTIVE. These 2 m sentences form the initial training set for the iterative learning methods.

4.2 Self-training method

Self-training first uses the initial training set to train a classifier using a standard supervised learning approach (NB classifier in our study) and then applies it to classify unlabeled sentences in the corpus. Based on their probabilities of being SUBJECTIVE or OBJECTIVE, n sentences with the highest confidence in each class

[3] Note that this is an empirical solution. We found that using log-based normalization yielded slightly better accuracy in self-labeling on MPQA and AMI data. We will evaluate other methods for creating the initial labeled set in the future.

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

Algorithm 1 Self-training.

Define U: unlabeled data set
Define T: training set
Create initial pseudo training set T_0 from U using lexicon-based method
$T = T_0$
$U = U - T_0$
while iteration continues **do**
  train classifier using T
  use trained classifier to label U
  S_i = top ranked n subjective and n objective examples
  $T = T + S_i$
  $U = U - S_i$
end while

are selected. They are added to the training set and removed from the unlabeled set. This process repeats until it reaches the predefined iteration number or uses all the unlabeled data. This self-training algorithm is described in Algorithm 1.

This self-training method has some drawbacks that may affect the performance. The instances added to the training set in each iteration are labeled by the current classifier and may not be accurate, because of the classifier's performance. Furthermore, the labels of the initial training set are derived from a rule-based classifier, which may be incorrect and thus impact the classifier trained from this data. However, once the automatically labeled instances are added to the training set, their labels do not have a chance to be corrected, even though new information may have been explored in later iterations and classifiers become more accurate.

4.3 Calibrated EM naive bayes method

The EM NB method is a semi-supervised algorithm proposed in Nigam *et al.* (2000) for learning from both labeled and unlabeled data and has been used in many text classification tasks (Dai *et al.* 2007; Druck *et al.* 2007). We expect this approach can address some of the problems in the self-training method mentioned above.

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 NB classifier to estimate the posterior probabilities of each sentence s_i belonging to each class c_j (SUBJECTIVE and OBJECTIVE), $P(c_j|s_i)$:

$$\begin{aligned} P(c_j|s_i) &= \frac{P(c_j)P(s_i|c_j)}{P(s_i)} \\ &= \frac{P(c_j)\prod_{k=1}^{|s_i|}P(w_k|c_j)}{\sum_{c_l\in C}P(c_l)\prod_{k=1}^{|s_i|}P(w_k|c_l)} \end{aligned} \tag{3}$$

The M-step uses the probabilistic results from the E-step to recalculate the parameters in the NB classifier as follows:

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

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

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

Equation (4) computes the probability of word w_t in class c_j. Equation (5) calculates the prior probability of class c_j. 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. That is, the classifier is built using the initial training data without any soft probabilities, just like in self-training.

There are some differences between self-training and EM. Self-training uses only the labeled sentences in the training set for NB classifier training, whereas EM uses the entire data set, with soft labels for each instance in each iteration (except the initial step). In addition, the labeling step is different between them: self-training uses the current classifier to classify those unlabeled sentences (the unlabeled data size reduces in every iteration); whereas EM relabels all the initial unlabeled data in every iteration. We expect that the soft labels in EM will be useful for learning if the posterior probabilities estimated by the classifier are reasonably accurate.

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 distribution 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 are 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, and the unlabeled data indeed have a different class distribution (e.g., due to data mismatch), then the distribution adjustment performed in calibrated EM may hurt system performance.

5 Empirical evaluation

In this section, we evaluate our unsupervised learning methods and analyze various impacting factors. For subjectivity classification experiments, we use classification

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

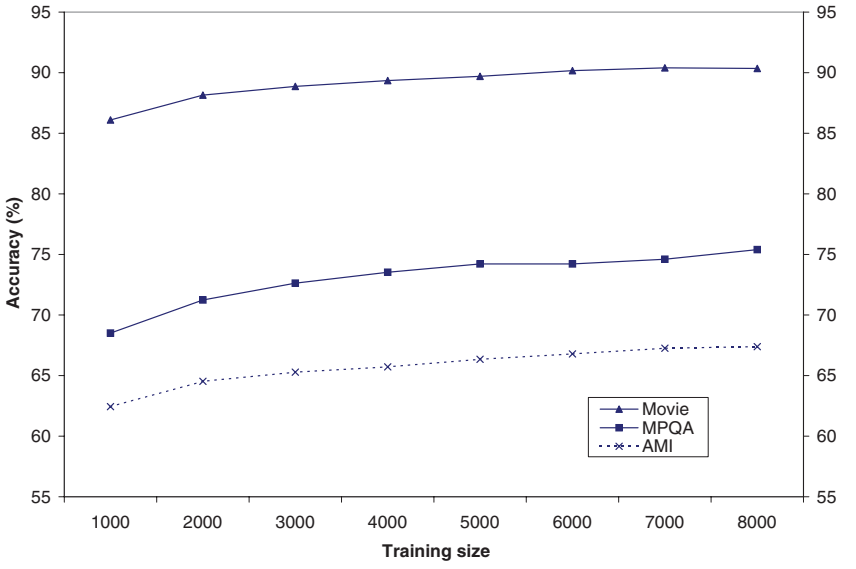


Fig. 1. Learning curves of the supervised NB classifier on movie, MPQA, and AMI data.

accuracy as performance metrics. In preprocessing, we removed the punctuation and numbers from the data and performed word stemming using the TreeTagger toolkit.[4] We did not remove the stopwords, because in preliminary experiments we found doing that resulted in slight performance degradation. For all the experiments, only unigrams are used as features in the classifiers (i.e., bag-of-words model).

5.1 Supervised baseline

As mentioned before, we use NB classifier as the base classifier in self-training and calibrated EM learning, therefore, we first present results using a supervised NB classifier as a baseline. For each corpus, we reserve 2,000 samples as a held-out test set, and vary the training size from 1,000 to 8,000 using the remaining data.[5] Both training and test sets are balanced between the two classes. Figure 1 shows the learning curves (classification accuracy) using NB for the three test sets when the training size varies. For each training size, we performed several runs using randomly chosen sentences to form a training set, always with balanced class distribution. The reported results are the average of multiple runs.

Not surprisingly, we can see that overall when training size increases, classification performance improves. This is observed for all the three corpora. We can also see there is a substantial difference of classification performance for these three data sets—results are the best for the movie corpus and worst for the AMI data. This is

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[5] The maximum training size in AMI data is slightly less than 8,000 (it is 7,892). When there is no ambiguity, we are going to use 8,000 to represent the data size for all the domains.

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

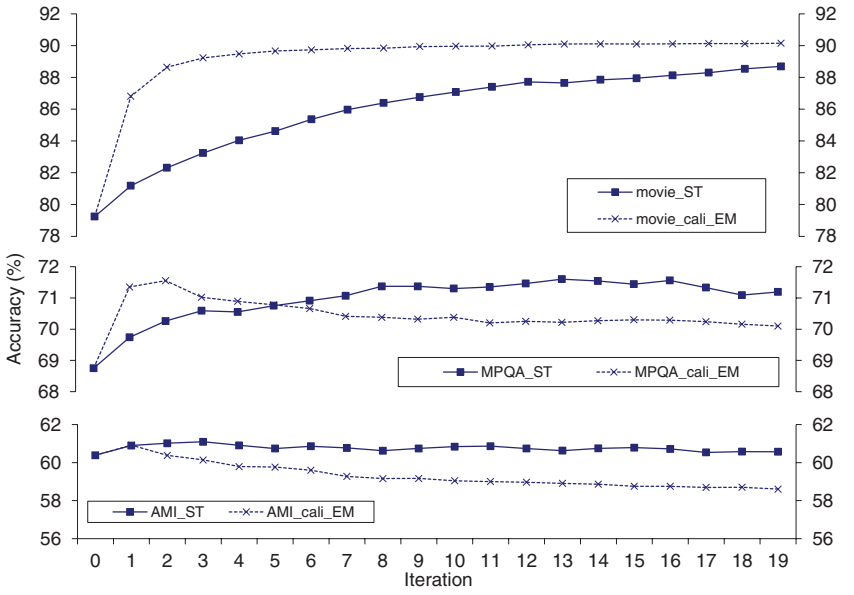


Fig. 2. Self-training (ST) and calibrated EM (cali_EM) results using unsupervised setting on movie, MPQA, and AMI data. The initial data set contains 2,000 self-labeled samples.

perhaps due to the genre difference among the three domains (we will show more error analysis in Section 5.3.4).

5.2 Unsupervised learning results

In unsupervised learning, we first use the lexicon-based method to generate the initial training set, then apply self-training and calibrated EM to incorporate unlabeled data in learning. In this set of experiments, we perform five-fold cross validation. We divide the corpus into five 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. For self-training, we add 100 instances from each class into the training set in each iteration. Figure 2 shows the learning curves of self-training and calibrated EM on movie, MPQA, and AMI data, respectively. Each point corresponds to the averaged five-fold cross-validation results. Classification performance for iteration 0 reflects the result of the classifier trained on the initial training set.

On movie data, both methods improve the performance significantly ($p < 0.001$ using paired t-test), compared to that based on the initial training set (iteration 0). In addition, calibrated EM outperforms self-training. It takes only a few iterations for the EM method to converge. At the end of the iteration, calibrated EM achieves 90.15% accuracy, which rivals the supervised learning performance of 91.31% when using all the 8,000 labeled sentences for training. On MPQA data, both methods

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

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

	movie			MPQA			AMI		
size	**sub**	**obj**	**Acc_Test**	**sub**	**obj**	**Acc_Test**	**sub**	**obj**	**Acc_Test**
100	95.20	82.20	59.93	83.20	87.60	60.45	49.60	71.60	50.51
200	90.20	82.00	71.63	85.60	86.60	63.83	53.40	71.00	53.81
1000	82.48	80.88	77.62	85.76	87.64	66.98	65.96	68.56	60.53
2000	79.24	79.04	79.24	85.18	87.46	68.75	66.98	69.04	60.39
3000	77.13	77.31	79.64	82.53	85.92	70.05	67.05	69.89	60.46

yield some improvement compared to the initial point[6] ($p < 0.05$). However, there is a different pattern between self-training and calibrated EM. Performance continues to increase with the iterations in self-training, while for EM we observe a peak accuracy in the first couple of iterations, but then performance starts dropping and eventually is worse than self-training. On AMI data, there is no performance improvement from either method. The result of self-training remains almost the same through the iterations, and calibrated EM degrades after the first iteration.

5.3 Analysis and discussion

We have observed a lot of differences among the three corpora in the above unsupervised learning results for self-training and calibrated EM. In this section, we conduct extensive analysis of impacting factors on the learning algorithms and the performance difference across domains.

5.3.1 Effect of initial set in unsupervised learning

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 of different initial size for the three domains. The initial set labeling accuracy is measured for SUBJECTIVE and OBJECTIVE class separately, i.e., the percentage of the correctly labeled sentences among the extracted sentences for that class (which is the widely used precision measure). In addition, we present the classification performance on the test set when using the NB classifier trained from the initial set (corresponding to performance at iteration 0 using iterative learning). Each size in the table represents the total number of sentences in the initial set (half are SUBJECTIVE sentences, and the other half are OBJECTIVE). Since we used

[6] In calibrated EM method, the comparison is conducted between the initial result and the peak point.

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

five-fold cross validation in the previous experiment, this results shown in the table are also the average of five runs.

Table 2 shows that when the size is 2,000 (as we used in previous experiments), the accuracy for both classes on MPQA is even better than on movies, even though we have seen that the 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 our rule-based method used to create the initial training set does not perform well on these two corpora because the most highly ranked sentences according to the subjective lexicon are not those most subjective sentences. We will see some examples in the error analysis, which shows that some objective sentences contain subjective lexicons.

From the results on the test set when using the initial self-labeled data, 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 examples. On AMI data, there is almost no difference among these 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 the test set can benefit more from using a larger initial training set, even though it is noisy (the accuracy of the self-labeled initial set often decreases when the set is larger).

In order to further investigate the impact of the self-labeled initial data set, we perform standard semi-supervised learning using reference labels in the initial data set. A similar setup as the unsupervised experiments in Section 5.2 is used here. We also perform five-fold cross validation, where for each fold we randomly select 2,000 samples (with their reference labels) as the initial training set from the pool of 8,000 samples. Figure 3 shows the learning curves of self-training and calibrated EM for the three corpora using semi-supervised learning.

On movie data, both methods yield 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. This shows that these semi-supervised learning approaches (especially calibrated EM) work well for this domain. On MPQA and AMI data, calibrated EM slightly increases the accuracy at the first iteration and then degrades thereafter. Self-training does not show much gain in either domain. For the three domains, the best results are achieved by calibrated EM. We notice that calibrated EM achieves a significantly higher accuracy after only one iteration. This shows that including all the unlabeled data in training (with

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

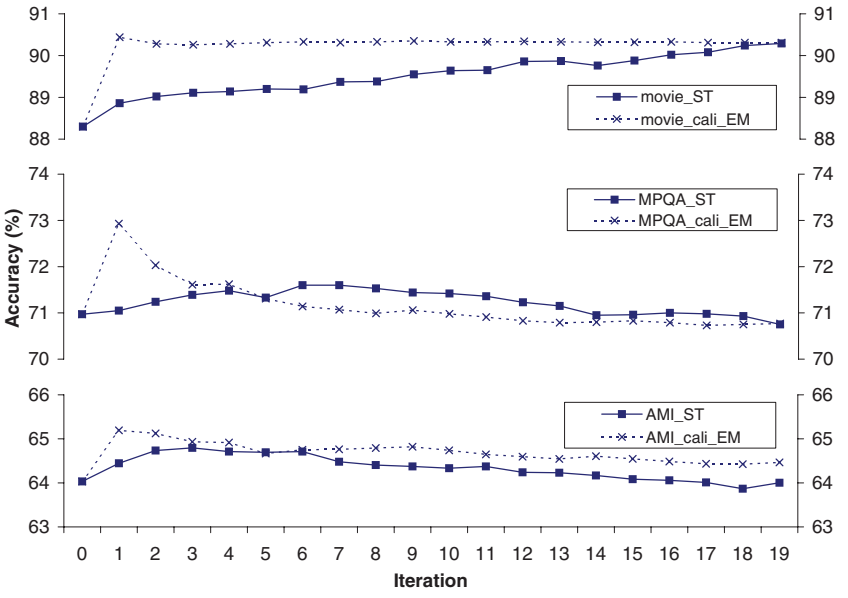


Fig. 3. Self-training (ST) and calibrated EM (cali_EM) results using semi-supervised learning (2,000 labeled seed) on movie, MPQA, and AMI data.

soft labels learned from the initial classifier) is helpful, however, more EM iterations do not yield further gain.

Comparing the results in unsupervised and semi-supervised learning (Figures 2 and 3), we find that on the movie domain, both self-training and calibrated EM methods show a similar pattern, even though there is some slight difference in the absolute performance (better performance in semi-supervised learning than the unsupervised one). On MPQA data, self-training does not yield much improvement in semi-supervised learning, which is surprising compared to the performance gain in unsupervised setting. On AMI data, self-training has a similar behavior in the two settings (no performance gain), and calibrated EM yields some improvement in the first iteration. The only difference between unsupervised and semi-supervised learning is the initial data set. Table 2 shows that the training set generated using the lexicon-based method has a very low accuracy on AMI data; however, this does not explain the poor performance of unsupervised iterative learning on this domain, because even when the initial set has 100% accuracy (i.e., semi-supervised setting), the two iterative learning methods fail to yield any performance gain on AMI data. We will analyze other factors that have an effect on learning performance in the following experiments.

5.3.2 *Impacting factors in self-training*

From the learning curves of unsupervised and semi-supervised learning, we notice that self-training continues to increase performance on movie data, but not always for the other data. There is a major difference among the three domains—we have

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

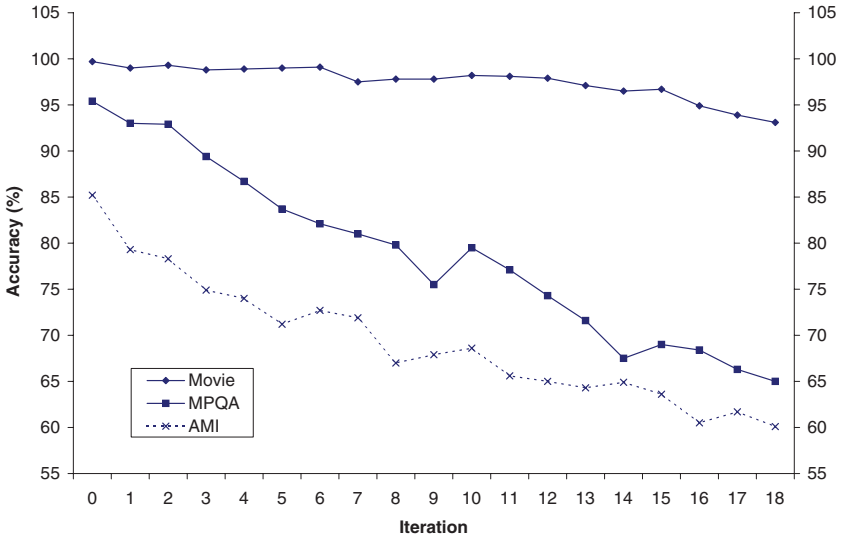


Fig. 4. Accuracy of added samples in each iteration in self-training for movie, MPQA, and AMI data.

seen from the supervised baseline results (in Figure 1), MPQA and AMI data are much harder to classify than movie data using the NB classifier. Thus, we infer that the poor classification capability of NB on MPQA and AMI data might be one reason to explain the different patterns of self-training in the three domains.

In self-training, the current classifier selects examples to add to the pool to train a new classifier in next iteration. Intuitively, these added examples have an impact on learning performance. Therefore, we first examine the accuracy of the added samples in each iteration. Figure 4 shows these results for the three domains. We use semi-supervised learning for this analysis, rather than unsupervised learning, in order to avoid the confounding impact of the initial training set accuracy. We can see that the accuracy of the added samples remains quite high in all the iterations on movie domain, but degrades a lot on other two domains. This is due to the low classification accuracy on these two domains—even though the selected samples have high posterior probabilities based on the current classifier, they are not reliable. This accuracy of the added examples partly explains why self-training performs much better on movie data than the other two domains.

To further evaluate the impact of the accuracy of the added samples, we conduct a controlled experiment in semi-supervised learning where in each iteration we add samples with controlled accuracy, instead of using hypothesized labels provided by the model. Note that we use the same samples as those added by the model in semi-supervised learning in order to eliminate the data effect and just focus on the accuracy of the labels. We find that those added samples are roughly balanced between the two classes (otherwise, it will change the class distribution in the training data). For each experiment, we randomly generate the same number of errors for each class. Results are presented in Figure 5 for MPQA and AMI data, with accuracy of 0.85, 0.9, and 0.95 for the added examples. We do not include

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

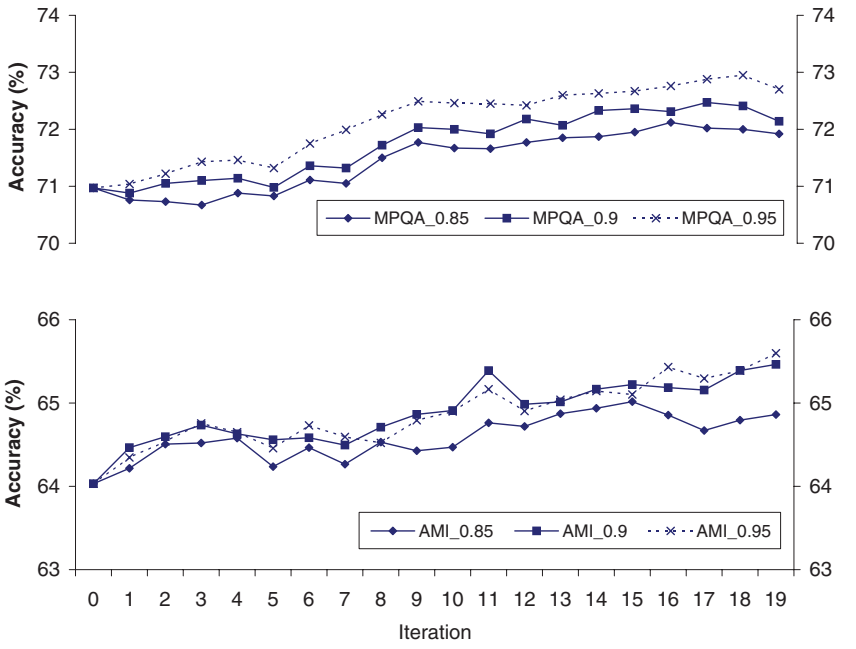


Fig. 5. Self-training performance using controlled accuracy for the added examples for MPQA and AMI data. MPQA_0.85 means that the accuracy of the labels for the added examples in each iteration is 0.85.

results for movie data in the figure since we have seen that it has a high self-labeling accuracy and self-training performs well. This experiment shows that when using a high accuracy in those added samples, self-training is indeed able to boost the performance. However, the low accuracy of the added samples in the real scenario limits the capability of self-training for MPQA and AMI data. This verifies our expectation.

An interesting observation from the unsupervised and semi-supervised learning results on MPQA data is that self-training is able to increase the accuracy from the noisy initial training set in the unsupervised setting (Figure 2), but in the semi-supervised one it degrades after some slight improvement (Figure 3). In addition, their peak points are similar, both around 71.5%. There seems to be a performance upper bound when using self-training. We believe this is because when the accuracy of the current classifier is high enough, the added inaccurate samples are too noisy to boost the performance; in contrast, when the current classifier has poor performance, the added noisy samples increase the training data size and can still yield some performance gain.

Another impacting factor in self-training is the initial training size. To analyze this, we perform semi-supervised learning experiments that use different initial training sizes and thus also different initial accuracies. The experiment results on MPQA and AMI data are shown in Figure 6 for data size of 200, 500, and 1,000 samples. It shows that on these two domains self-training only works when the initial labeled training set is small. That is, when the current classifier's performance is poor, it can

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

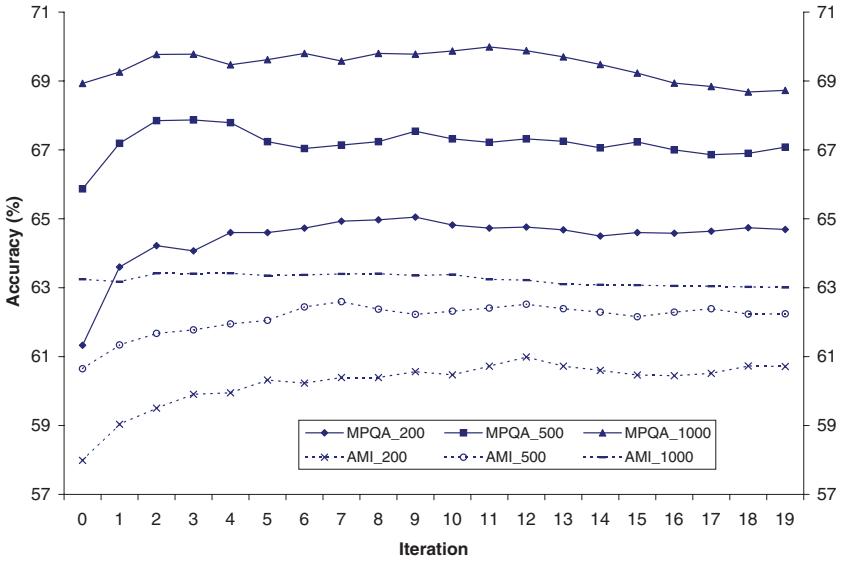


Fig. 6. Self-training with different labeled size on MPQA and AMI data. MPQA_200 means that the initial labeled training size is 200.

benefit from the added samples, but when the classifier performs well, the added incorrect samples cannot yield further improvement. This experiment also explains the different patterns on MPQA mentioned above when the initial data set has reference labels in semi-supervised learning versus automatically generated labels in unsupervised learning, respectively. We are not showing results for the movie data here since self-training consistently yields high accuracy using different initial data size. That is because on the movie data the NB classifier has a high accuracy and the selected samples from unlabeled data in each iteration are mostly correct and informative. We also evaluated the effect of the initial data size on unsupervised setting and found that, in general, the patterns are similar to semi-supervised learning, but with lower absolute performance.

5.3.3 Issues related to EM

Similar to self-training, we also examine the effect of using different initial size in calibrated EM. Figure 7 shows the results for the three corpora using semi-supervised learning with calibrated EM learning. We also include the initial results (i.e., supervised learning using the initial training set) in the figure for a comparison. Different from the figures of the learning curves shown in early experiments, here each point in the graph represents the final result after the iterative learning process. In these experiments, we stop the iterations when the model probability converges (difference of the likelihood of the data in two iterations is less than a predefined threshold). We observe that on all the three domains, for a particular experimental setting the model probability converges very fast after increasing dramatically in the first several iterations. When the labeled data size increases, the change of model

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

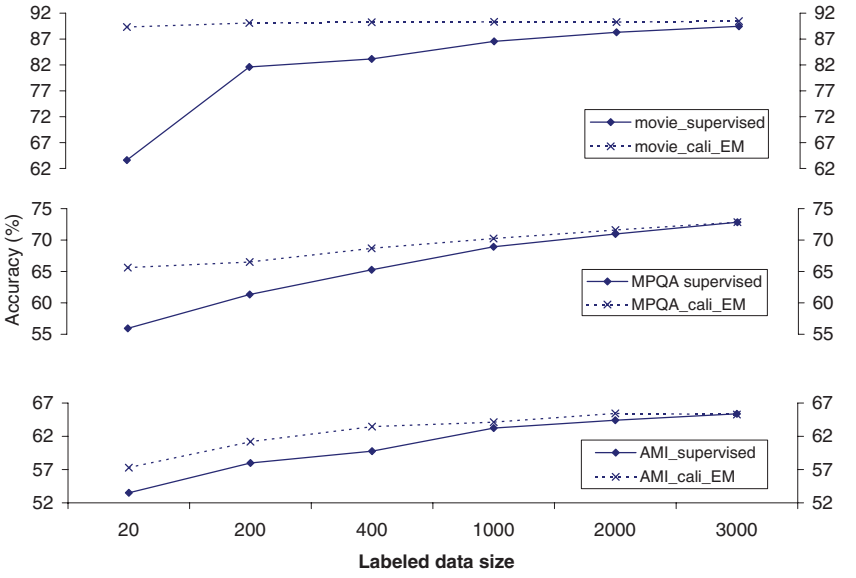


Fig. 7. Comparison of calibrated EM and supervised baseline using different initial training size.

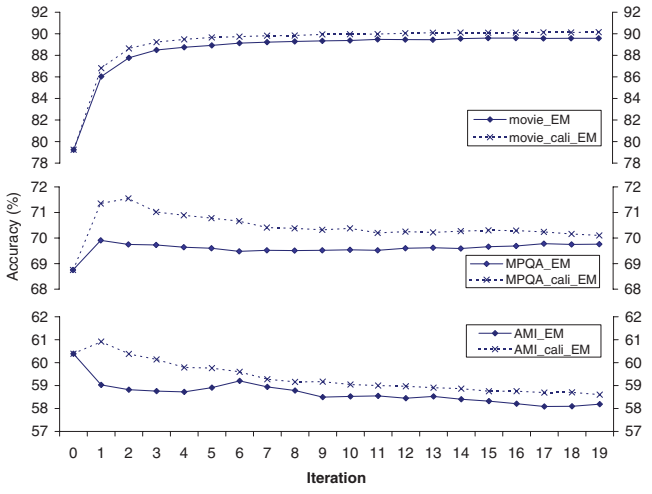
probability is not as much as using smaller data size, which is as expected: when the labeled size is large enough, unlabeled data does not provide much additional information.

On movie data, it is surprising to see that when given only 20 labeled samples, calibrated EM can boost the accuracy to 89%, which is as high as using training data with 3,000 labeled samples in supervised learning. When there are more initial samples (e.g., 2,000), calibrated EM still provides some performance gain. While on MPQA and AMI data, calibrated EM increases the accuracy only when the labeled data set is small and the original performance is poor. When given 2,000 labeled samples, there is a marginal performance increase. When the initial data set increases further (e.g., 3,000 samples), we can see performance even degrades a bit compared to the initial result. This is what we observed earlier in Figure 3—we found that calibrated EM performs very well on movie data, but on both MPQA and AMI data, it decreases the accuracy after it achieves the highest accuracy (always at the first or second iteration), instead of converging to the peak point.

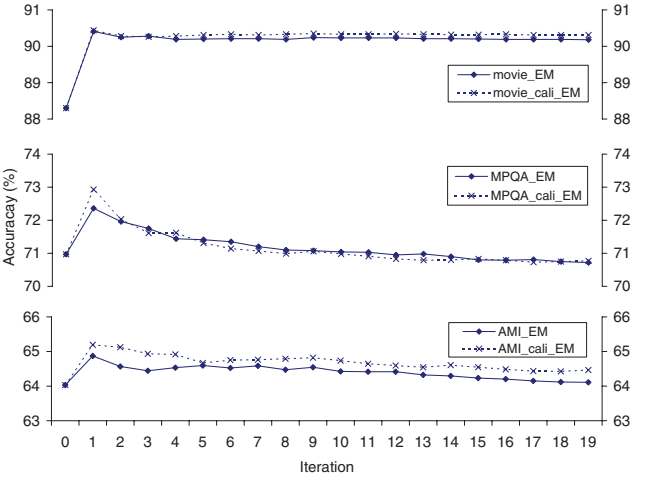
We believe this is related to the data property and the assumptions used in EM. Similar patterns have been found in some previous studies (Chapelle, Schölkopf and Zien 2006). They attribute this to the incorrect model assumption, i.e., when the modeling assumptions for a particular classifier do not match the characteristics of the distribution of the data, unlabeled data may degrade the performance of the classifiers (Chapelle *et al.* 2006). The low performance of NB on MPQA and AMI data in supervised results (Figure 1) may also indicate the mismatch of the model and data distribution.

Another point related to EM is the calibrated EM method used in this paper. Figure 8 compares calibrated EM with standard EM on the three domains using

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press



(a) unsupervised



(b) semi-supervised

Fig. 8. Comparison of standard EM and calibrated EM.

unsupervised (a) and semi-supervised learning (b), respectively. We can see that calibrated EM outperforms standard EM, with a larger improvement on MPQA and AMI data, especially in unsupervised learning. When using standard EM in unsupervised learning, 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 (which is more balanced). 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 MPQA and AMI data because it adjusts the posterior probability of each sample according to

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

the class distribution in the data, making it more accurate in training the model in the next iteration.

5.3.4 Discussion

There are a few points worth discussing based on our experiments and error analysis.

A. **Class distribution.**

All of our experiments are performed on balanced data sets since these data have been widely used in previous research. A natural question to ask is if this data distribution has any effect on our unsupervised learning methods. For an analysis, we applied the same methods on imbalanced data sets, using a ratio of 1:3 for SUBJECTIVE to OBJECTIVE as an example. The initial labeled data and added samples in self-training are modified to have the same distribution as in the whole data set. For calibrated EM, we also adjust the posterior probabilities accordingly. We found that the results on the imbalanced data have similar patterns as those observed on the balanced data. Both unsupervised iterative learning methods work well on movie and MPQA data, but not on AMI data. Note that in our study, we assume we know the data distribution *a prior*. In real testing, this information may not be available. Some development sets will be needed to optimize the classifier's decision threshold in order to take into account different class distributions for training and testing. This is a general problem for many machine learning problems, not specific to our methods in this work.

B. **Domain difference.**

Much of the difference we have observed (classification accuracy, effectiveness of iterative learning) can be attributed to the genre difference. In movie reviews (subjective sentences in movie data), often a person expresses his/her favor (or not) of the movie explicitly. From Table 1, we can see that the vocabulary size is larger in this data set. The diverse words cause less ambiguity and correspond to more indicative features for classification. In addition, these data were created by collecting sentences from two different sources: movie reviews and plots. These explain why subjectivity detection is relatively easier in this data set.

MPQA data are collected from news resources, where 'subjectivity' mostly means an opinion, an attitude, or a judgment. Take 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 is ambiguous as 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. The small vocabulary in this data results in some ambiguity. There are many incomplete sentences and disfluencies (repeated or revised segments). 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',

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

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 from the same person in conversations.

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

We notice that many of the classification errors in AMI occur in very short sentences, such as in the example shown above. Using the calibrated EM unsupervised learning output, we found that the accuracy for utterances with less than five words is 56.07%, and 62.47% for longer utterances, which shows that short sentences are very ambiguous for subjectivity classification. As Table 1 shows, the sentence length is much shorter in AMI than the other two corpora, which is one reason for the poor performance on the AMI data.

We also find that on movie data, the accuracy for the two classes is similar; however, on AMI and MPQA data, the system has a high recall but low precision for SUBJECTIVE class, and low recall but high precision for OBJECTIVE class. This is the case for all the settings: both self-training and calibrated EM learning methods, and both semi-supervised and unsupervised settings. The system tends to classify more objective samples as subjective in MPQA and AMI data. This observation is also related to the effectiveness of calibrated EM over standard EM discussed before—it is more helpful for AMI and MPQA because of the different performance for the two classes, thus unbalanced data distribution in its labeling results.

C. **Limitation of the current model.**

Our error analysis 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** 1950s housewife, living the **perfect** 1950s 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

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

several subjective words (shown in bold) and are therefore labeled as subjective. These are the problems with the current word based approaches, both lexicon-based initial labeling or supervised classifiers using bag-of-words models.

There has been some previous work using an expanded subjective lexicon. Wiebe and Riliff (2005) extracted new lexico-syntactic patterns from unannotated data. Andreevskaia and Bergler (2008) incorporated two systems, one trained with in-domain data and the other trained on a general dictionary wordnet. Choi and Cardie (2009) adapted a general polarity lexicon into a domain-specific one using the relation between words in the same expression. In addition, a lot of work has investigated features other than bag-of-words, such as part-of-speech tags and dependency relations. Their results show that using a larger lexicon or rich feature sets will yield some improvement in subjectivity identification, but often the improvement is rather limited and may not generalize to different domains. In this study, we choose to use the lexicon as in Wilson et al. (2005), the bag-of-words model, and a base NB classifier, since our focus is on a cross-corpus evaluation of two unsupervised learning methods, and these baseline methods are quite competitive and robust to different domains. While using an expanded lexicon and additional features may slightly improve the absolute classification performance, we expect that the general patterns about the unsupervised learning methods and the impact factors will remain the same. We will evaluate these in our future work.

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 iterative learning methods, including self-training and calibrated EM. More importantly, in order to study the genre effect, we evaluate these methods across three different data sets, one being meeting dialog that has not been widely used in sentiment analysis. Significant difference has been observed across these corpora in terms of the effectiveness of these learning approaches. Both bootstrapping methods yield good performance on movie data and can achieve performance similar to using fully supervised learning. However, for MPQA corpus, the performance gain is not as much as on movie data, while on AMI corpus both methods fail to yield improvement. In general, the calibrated EM method converges faster and have a better performance than self-training. We analyzed the effect of the automatically created initial data set, compared it to semi-supervised learning where a labeled initial set is used, and examined various factors in self-training and calibrated EM. Our analysis showed that the performance of the base classifier has a substantial impact on iterative learning methods, which explained the different learning patterns for different corpora. For corpora with low classification accuracy, the learning methods are useful only when the initial training size is small and initial accuracy is low. In addition, we found that calibrated EM outperforms standard EM when the class distribution based on classifier's hypotheses does not match that in

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

the data set. Our error analysis also showed that the current word-based approaches have limitations for subjectivity detection, especially in MPQA and AMI data.

Our iterative learning approaches use a base NB classifier. One of the problems with NB classifiers is that their posterior probabilities are not accurate. Therefore, in our future work, we will evaluate using other base models. We will also investigate other semi-supervised learning methods to utilize unlabeled data and improve self-labeling to create more accurate initial training set. 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.

References

Andreevskaia, A., and Bergler, S. 2008. When specialists and generalists work together: overcoming domain dependence in sentiment tagging. In *Proceedings of ACL/HLT*, Columbus, Ohio.

Chapelle, O., Schölkopf, B., and Zien, A. (eds). 2006. *Semi-Supervised Learning*. Cambridge, MA: MIT Press.

Choi, Y., and Cardie, C. 2009. Adapting a polarity lexicon using integer linear programming for domainspecific sentiment classification. In *Proceedings of EMNLP*, Singapore.

Dai, W., Xue, G.-R., Yang, Q., and Yu, Y. 2007. Transferring naive Bayes classifiers for text classification. In *Proceedings of AAAI*, Vancouver, British Columbia, Canada.

Dasgupta, S., and Ng, V. 2009. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In *Proceedings of ACL-IJCNLP*, Suntec, Singapore.

Druck, G., Pal, C., McCallum, A., and Zhu, X. 2007. Semi-supervised classification with hybrid generative/discriminative methods. In *Proceedings of ACM SIGKDD*, San Jose, CA, USA.

Gyamfi, Y., Wiebe, J., Mihalcea, R., and Akkaya, C. 2009. Integrating knowledge for subjectivity sense labeling. In *Proceedings of NAACL*, Boulder, CO, USA.

Hu, M., and Liu, B. 2006. Opinion extraction and summarization on the web. In *Proceedings of AAAI*, Boston, MA, USA.

Kim, S.-M., and Hovy, E. 2005. Automatic detection of opinion bearing words and sentences. In *Proceedings of ACL*, Jeju Island, Korea.

Li, S., Huang, C.-R., Zhou, G., and Lee, S. Y. M. 2010. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *Proceedings of ACL*, Uppsala, Sweden.

Melville, P., Gryc, W., and Lawrence, R. D. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of ACM SIGKDD*, Paris, France.

Murray, G., and Carenini, G. 2008. Summarizing spoken and written conversations. In *Proceedings of EMNLP*, Honolulu, Hawaii.

Murray, G., and Carenini, G. 2009. Detecting subjectivity in multiparty speech. In *Proceedings of Interspeech*, Brighton, UK.

Nakagawa, T., Inui, K., and Kurohashi, S. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In *Proceedings of NAACL*, Los Angeles, CA, USA.

Ng, V., Dasgupta, S., and Arifin, S. M. N. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of COLING/ACL*, Sydney, Australia.

Ni, X., Xue, G.-R., Ling, X., Yu, Y., and Yang, Q. 2007. Exploring in the weblog space by detecting informative and affective articles. In *Proceedings of WWW*, Banff, Alberta, Canada.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* **39**: 103–34.

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press

Nishikawa, H., Hasegawa, T., Matsuo, Y., and Kikui, G. 2010. Optimizing informativeness and readability for sentiment summarization. In *Proceedings of ACL*, Uppsala, Sweden.
Pang, B., and Lee, L. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*, Barcelona, Spain.
Pang, B., and Lee, L. 2008. Using very simple statistics for review search: An exploration. In *Proceedings of COLING*, Manchester, UK.
Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, Philadelphia, PA, USA.
Raaijmakers, S., and Kraaij, W. 2008. A shallow approach to subjectivity classification. In *Proceedings of ICWSM*, Seattle, DC, USA.
Raaijmakers, S., Truong, K., and Wilson, T. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of EMNLP*, Honolulu, Hawaii.
Riloff, E., and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP*, Stroudsburg, PA, USA.
Riloff, E., Wiebe, J., and Phillips, W. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of AAAI*, Pittsburgh, PA, USA.
Sebastiani, F., Esuli, A., and Sebastiani, F. 2006. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of EACL*, Trento, Italy.
Tsuruoka, Y., and Tsujii, J. 2003. Training a naive Bayes classifier via the EM algorithm with a class distribution constraint. In *Proceedings of NAACL*, Edmonton, Canada.
Wiebe, J., and Riloff, E. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing*, Mexico City, Mexico.
Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. 2004. Learning subjective language. *Computational Linguistics* **30**(3): 277–308.
Wiegand, M., and Klakow, D. 2010. Bootstrapping supervised machine-learning polarity classifiers with rule-based classification. In *Proceedings of WASSA*, Lisbon, Portugal.
Wilson, T. 2008. Annotating subjective content in meetings. In *Proceedings of LREC*, Marrakech, Morocco.
Wilson, T., and Wiebe, J. 2003. Annotating opinions in the world press. In *Proceedings of SIGdial*, Sapporo, Japan.
Wilson, T., Wiebe, J., and Hwa, R. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI*, San Jose, CA, USA.
Wilson, T., Wiebe, J., and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, Vancouver, British Columbia, Canada.
Yu, H., and Hatzivassiloglou, V. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, Stroudsburg, PA, USA.
Zhou, S., Chen, Q., and Wang, X. 2010. Active deep networks for semi-supervised sentiment classification. In *Proceedings of COLING*, Beijing, China.

https://doi.org/10.1017/S1351324911000234 Published online by Cambridge University Press