Filling the Gap:
Semi-Supervised Learning for Opinion Detection Across Domains

Abstract

We investigate the use of Semi-Supervised Learning (SSL) in opinion detection both in sparse data situations and for domain adaptation. We show that co-training reaches the best results in an in-domain setting with small labeled data sets, with a maximum absolute gain of 33.5%. For domain transfer, we show that self-training gains an absolute improvement in labeling accuracy for blog data of 16% over the supervised approach with target domain training data.

1 Introduction

Rich and free opinions published electronically and, more recently, on the WWW offer ample opportunities to discover individual's attitudes towards certain topics, products, or services. To capitalize on this enormous body of opinions, researchers have been working in the area of opinion mining since the late 1990s. Opinion detection seeks to automatically determine the presence or absence of opinions in a text, and it is therefore a fundamental task for opinion mining.

In order to capture subtle and creative opinions, opinion detection systems generally assume that a large body of opinion-labeled data are available. However, collections of opinion-labeled data are often limited, especially at the granularity level of sentences; and manual annotation is tedious, expensive and error-prone. The shortage of opinion-labeled data is less challenging in some data domains (e.g., reviews) than in others (e.g., blog posts). A simple method for improving accuracies in challenging domains would be to borrow opinion-labeled data from a non-target data domain; but this approach often fails because opinion detection strategies designed for one data domain generally do not perform well in another domain. One reason for failure of the simple transfer approach is that the information used for opinion detection is typically lexical, and lexical means of expressing opinions may vary not only from domain to domain, but also from register to register. For example, while the word ”awesome” is a good indicator of an opinion in blogs, it is less likely to occur in the same role in newspaper texts.

While it is difficult to obtain opinion-labeled data, one can easily collect almost infinite unlabeled user-generated data that contain opinions. The use of Semi-Supervised Learning (SSL), motivated by limited labeled data and plentiful unlabeled data in the real world, has achieved promising results in various NLP studies (e.g., (Fürstenau and Lapata, 2009; Talukdar and Pereira, 2010)), yet it has not been fully investigated for use in opinion detection. Although studies have shown that simple SSL methods are promising for extracting opinion features or patterns using limited opinion-labeled data (e.g., (Wiebe and Riloff, 2005)), few efforts have been made either to apply SSL directly to opinion detection or to examine more sophisticated SSL methods. This research is intended to fill the gap regarding application of SSL in opinion detection. We investigate a range of SSL algorithms with a focus on self-training and co-training in three types of electronic documents: edited news articles, semi-structured movie reviews, and the informal and unstructured content of the blogosphere. We conclude that SSL is a successful method for handling the shortage of opinion labeled data and the domain transfer problem.

2 Background and Related Work

There is a wide range of literature on opinion detection. We concentrate here on supervised and semi-supervised approaches.

2.1 Supervised Learning for Opinion Detection

Supervised learning algorithms that can automatically learn important opinion-bearing features from an annotated corpus have been adopted and investigated for opinion detection and yielded satisfying results (Wiebe et al., 2004; Yu and Hatzivassiloglou, 2003; Zhang and Yu, 2007). With no classification techniques developed specifically for opinion detection, state-of-the-art topical supervised classification algorithms can achieve performance comparable to complex linguistic approaches when using binary values (i.e., presence or absence) and incorporating different types of features. Commonly used opinion-bearing features include bag-of-words, POS tags, ngrams, low frequency words or unique words (Wiebe et al., 2004; Yang et al., 2007), semantically oriented adjectives (e.g., “great”, ”poor”) and more complex linguistic patterns. Both the scale and quality of the annotated corpus play an important role in the supervised learning approach.

2.2 SSL and Opinion Detection Applications

In contrast to supervised learning, SSL learns from both labeled and unlabeled data. SSL assumes that, although unlabeled data hold no information about classes (e.g., “opinion” or “non-opinion”), they do contain information about joint distribution over classification features. Therefore, when a limited set of labeled data is available in the target domain, using SSL with unlabeled data is expected to achieve an improvement over supervised learning.

Self-training Self-training is the simplest and most commonly adopted form of SSL for opinion detection. Self-training was originally used to facilitate automatic identification of opinion-bearing features. For example, Riloff and Wiebe (2003) proposed a bootstrapping process to automatically identify subjective patterns. Self-training has also been applied directly for identifying subjective sentences by following a standard self-training procedure: (1) train an initial supervised classifier on the labeled data; (2) apply this classifier to unlabeled data and select the most confidently labeled data, as determined by the classifier, to augment the labeled data set; and (3) re-train the classifier by restarting the whole process. Wiebe and Riloff (2005) used a self-trained Naïve Bayes classifier for classifying subjective sentences and achieved better recall with modest precision over several rule-based classifiers.

One shortcoming of self-training is that the resulting data may be biased: That is, the final labeled data may consist of examples that are easiest for this particular opinion detector to identify.

Co-training The core idea of co-training is to use two classifiers and trade additional examples between them, assuming that the resulting union of classified examples is more balanced than examples resulting from using either classifier alone. When labeling new examples, a final decision is made by combining the predictions of the two updated learners. The original co-training algorithm assumes redundancy in the training data and thus more than one view can be used to represent and classify each example independently and successfully (Blum and Mitchell, 1998). For example, an image can be naturally represented by its text description or by its visual attributes. Even when a natural split in the feature set is not available, studies have shown that the key to co-training is the existence of two largely different initial learners, regardless of whether they are built by using two feature sets or two learning algorithms (Wang and Zhou, 2007).

When there are different views for the target examples, co-training is conceptually clearer than self-training, which simply mixes features. Since co-training uses each labeled example twice, it requires less labeled data and converges faster than self-training. However, the lack of natural feature splits has kept researchers from exploring co-training for opinion detection. To the best of our knowledge, the only co-training application for opinion detection was reported by Jin et al. (2009), who created disjoint training sets for building two initial classifiers and successfully identified opinion sentences in camera reviews by selecting auto-labeled sentences agreed upon by both classifiers.

EM-Based SSL Expectation-Maximization (EM) refers to a class of iterative algorithms for maximum-likelihood estimation when dealing with incomplete data. Nigam et al. (1999) combined EM with a Naïve Bayes classifier to resolve the problem of topical classification, where unlabeled data were treated as incomplete data. The EM-NB SSL algorithm yielded better performance than either an unsupervised lexicon-based approach or a supervised approach for sentiment classification in different data domains, including blog data (Aue and Gamon, 2005; Takamura et al., 2006). No opinion detection applications of EM-based SSL have been reported in the literature.

S^{3}VMs Semi-Supervised Support Vector Machines (S^{3}VMs) are a natural extension of SVMs in the semi-supervised spectrum. They are designed to find the maximal margin decision boundary in a vector space containing both labeled and unlabeled examples. Although SVMs are the most favored supervised learning method for opinion detection, S^{3}VMs have not been used in opinion detection.

Graph-based SSL learning has been successfully applied for opinion detection (Pang and Lee, 2004), but it is not appropriate for dealing with large scale data sets.

2.3 Domain Adaptation for Opinion Detection

When there are few opinion-labeled data in the target domain and/or when the characteristics of the target domain make it challenging to detect opinions, opinion detection systems usually borrow opinion-labeled data from other data domains. This is especially common in opinion detection in the blogosphere (Chesley et al., 2006). To evaluate this shallow approach, Aue and Gamon (2005) compared four strategies for utilizing opinion-labeled data from one or more non-target domains and concluded that using non-targeted labeled data without an adaptation strategy is less efficient than using labeled data from the target domain, even when the majority of labels are assigned automatically by a self-training algorithm.

Blitzer et al. (2007) and Tan et al. (2009) implemented domain adaptation strategies for sentiment analysis. Although promising, their domain adaptation strategies involved sophisticated and computationally expensive methods for selecting general features to link target and non-target domains.

3 Motivation and Objective

While SSL is especially attractive for opinion detection because it only requires a small number of labeled examples, the studies described in the previous section have concentrated on simple SSL methods. We intend to fill this research gap by comparing the feasibility and effectiveness of a range of SSL approaches for opinion detection. Specifically, we aim to achieve the following goals:

First, to gain a more comprehensive understanding of the utility of SSL in opinion detection. We examine four major SSL methods: self-training, co-training, EM-NB, and S^{3}VM. We focus on self-training and co-training because they are both wrapper approaches that can be easily adopted by any existing opinion detection system.

Second, to design and evaluate co-training strategies for opinion detection. Since recent work has shown that co-training is not restricted by the original multi-view assumption for target data and that it is more robust than self-training, we evaluate new co-training strategies for opinion detection.

Third, to approach domain transfer using SSL, assuming that SSL can overcome the problem of domain-specific features by gradually introducing targeted data and thus diminishing bias from the non-target data set.

4 SSL Experiments

Our research treats opinion detection as a binary classification problem with two categories: subjective sentences and objective sentences. It is evaluated in terms of classification accuracy.

Since a document is normally a mixture of facts and opinions (Wiebe et al., 2001), sub-document level opinion detection is more useful and meaningful than document-level opinion detection. Thus, we conduct all experiments on the sentence level.

The remainder of this section explains the data sets and tools used in this study and presents the experimental design and parameter settings.

4.1 Data Sets

Three types of data sets have been explored in opinion detection studies: news articles, online reviews, and online discourse in blogs or discussion forums. These three types of text differ from one another in terms of structure, text genre (e.g., level of formality), and proportion of opinions found therein. We selected a data set from each type in order to investigate the robustness and adaptability of SSL algorithms for opinion detection and to test the feasibility of SSL for domain adaptation.

Movie Review One of the standard data sets in opinion detection is the movie review data set created by Pang and Lee (2004). It contains 5,000 subjective sentences or snippets from the Rotten Tomatoes pages and 5,000 objective sentences or snippets from IMDB plot summaries, all in lowercase. Sentences containing less than 10 tokens were excluded and the data set was labeled automatically by assuming opinion inheritance: every sentence in an opinion-bearing document expresses an opinion, and every sentence in a factual document is factual. Although this assumption appears to be acceptable for movie review data, it is generally unreliable for other domains.

News Article The Wall Street Journal part of the Penn Treebank III has been manually augmented with opinion related annotations. This set is widely used as a gold-standard corpus in opinion detection research. According to the coding manual (Wiebe et al., 1999), subjective sentences are those expressing evaluations, opinions, emotions, and speculations. For our research, 5,174 objective sentences and 5,297 subjective sentences were selected based on the absence or presence of manually labeled subjective expressions.

JDPA Blog Post The JDPA corpus (Kessler et al., 2010) is a new opinion corpus released in 2010. It consists of blog posts that express opinions about automobile and digital cameras with named entities and sentiments expressed about them manually annotated. For our purpose, we extracted all sentences containing sentiment-bearing expressions as subjective sentences and manually chose objective sentences from the rest by eliminating subjective sentences that were not targeted to any labeled entities. After this process, we had approximately 10,000 subjective sentences and 4,348 objective sentences. To balance the number of subjective and objective sentences, we used 4,348 sentences from each category.

4.2 Data Preparation

We removed a small number of stop words. No stemming was conducted since the literature shows no clear gain from stemming in opinion detection. One reason for this may be that stemming actually erases subtle opinion cues such as past tense verbs. All words were converted to lowercase and numbers were replaced by a placeholder #. Both unigrams and bigrams were generated for each sentence.

Each data set was randomly split into three portions: 5% of the sentences were selected as the evaluation set and were not available during SSL and supervised learning (SL) runs; 90% were treated as unlabeled data (U) for SSL runs and i% ($1 \leq i \leq 5$) as labeled data (L) for both SL and SSL runs. For each SSL run, a baseline SL run was designed with the same number of labeled sentences (i%) and a full SL run was designed with all available sentences (90% + i%). If effective, an SSL run would significantly outperform its corresponding baseline SL run and approach the performance of full SL run.

4.3 Experimental Design

We conducted three groups of experiments: 1) to investigate the effectiveness of the SSL approach for opinion detection; 2) to explore different co-training strategies; and 3) to evaluate the applicability of SSL for domain adaptation.

4.3.1 General Settings for SSL

The Naïve Bayes classifier was selected as the initial classifier for self-training because of its ability to produce prediction scores and to work well with small labeled data sets. We used binary values for unigram and bigram features, motivated by the brevity of the text unit at the sentence level as well as by the characteristics of opinion detection, where occurrence frequency has proven to be less influential. We implemented two feature selection options: Chi square and Information Gain.

Parameters for SSL included: (1) Threshold k for number of iterations. If k is set to 0, the stopping criterion is convergence; (2) Number of unlabeled sentences available in each iteration u ($u << U$); (3) Number of opinion and non-opinion sentences, p and n, to augment L during each iteration; and (4) Weighting parameter λ for auto-labeled data. When λ is set to 0, auto-labeled and labeled data are treated equally; when λ is set to 1, feature values in an auto-labeled sentence are multiplied by the prediction score assigned to the sentence.

We used the WEKA data mining software (Hall et al., 2009) for data processing and classification of the self-training and co-training experiments. EM implemented in LingPipe (Alias-i, 2008) was used for the EM-NB runs. S^{3}VMs implemented in SVM^{light} (Joachims, 1999) and based on local search were adopted for the S^{3}VM runs. Since hyper-parameter optimization for EM-NB and S^{3}VM is not the focus of this research and preliminary explorations on parameter settings suggested no significant benefit, default settings were applied for EM-NB and S^{3}VM.

4.3.2 Co-Training Strategies

For co-training, we investigated five strategies for creating two initial classifiers following the criteria that these two classifiers either capture different features or based on different learning assumptions.

Two initial classifiers were generated: (1) Using unigrams and bigrams respectively to create two classifiers based on the assumption that low-order n-grams and high-order n-grams contain redundant information and represent different views of an example: content and context; (2) Randomly splitting feature set into two; (3) Randomly splitting training set into two; (4) Applying two different learning algorithms (i.e., Naïve Bayes and SVM) with different biases; and (5) Applying a character-based language model and a bag-of-words model where the former takes into consideration the sequence of words while the latter does not. In practice, for strategy (1), bigrams were used in combination with unigrams because bigrams alone are weak features when extracted from limited labeled data at sentence level.

Auto-labeled sentences were selected if they were assigned a label that both classifiers agreed on with highest confidence. Because our initial classifiers violated the original co-training assumptions, forcing agreement between confident predictions improved the maintenance of high precision.

4.3.3 Self-Training for Domain Adaptation

Based on the literature and our preliminary results, movie reviews would achieve the highest accuracy while news articles and blog reviews would be considerably more challenging. Thus, we decided to use movie reviews as source data while news articles and blog posts were treated as target data domains. While the data split for the target domain remains the same as in section 4.2, all sentences in the source domain, except for the 5% evaluation data, were treated as labeled data. For example, in order to identify opinion-bearing sentences from the blog data set, all 9,500 movie review sentences and i% of blog sentences were used as labeled data, 90% of blog sentences were used as unlabeled data, and 5% as evaluation data. We also applies a parameter to gradually decrease the weight of the source domain data, similar to the work done by Tan et al. (2009).

	# Labeled Examples					
Type	100	200	300	400	500	all
Self-tr	85.2	86.6	87.0	87.2	86.6	
SL	63.8	73.6	77.2	79.4	80.2	89.4
Co-tr.	92.2	93.8	92.6	93.2	91.4	
SL	75.8	80.8	82.6	85.2	84.8	95.2
EM-NB	88.1	88.7	88.6	88.4	89.0	
SL	73.5	78.7	81.3	82.8	83.9	91.6
S^{3}VM	59.0	68.4	67.8	67.0	75.2	
SL	70.0	72.8	75.6	76.2	80.0	90.0

Table 1: Classification accuracy(%) of SSL and SL on movie reviews

5 Results and Evaluation

Overall, our results suggest that SSL improves accuracy for opinion detection although the contribution of SSL varies across data domains and different strategies need to be applied to achieve optimized performance. For the movie review data set, almost all SSL runs outperformed their corresponding baseline SL runs and approached full SL runs; for the news article data set, SSL performance followed a similar trend but with only a small rate of increase; for the blog post data set, SSL runs using only blog data showed no benefits over the SL baseline, but with labeled movie review data, SSL runs produced results comparable with full SL result.

5.1 SSL vs. SL

Table 1 reports the performance of SSL and SL runs on movie review data based on different numbers of initial labeled sentences. Both the self- and co-training runs reported here used the same parameter settings: k=0, u=20, p=2, n=2, λ =0, with no feature selection. The co-training results in Table 1 used a character-based language model and a bag-of-words model (see section 5.2). SL runs for co-training classified sentences based on the highest score generated by two classifiers; SL runs for S^{3}VM applied the default SVM setting in SVM^{light}; and SL runs for EM-NB used the Naïve Bayes classifier in the EM-NB implementation in LingPipe.

Table 1 shows that, except for S^{3}VM, SSL always outperforms the corresponding SL baseline on movie reviews: When SSL converges, it achieves improvement in the range of 8% to 34% over the SL baseline. The fewer initial labeled data, the more benefits an SSL run gained from using unlabeled data. For example, using 100 labeled sentences, self-training achieved a classification accuracy of 85.2% and outperformed the baseline SL by 33.5%. Although this SSL run was surpassed by 4.9% by the full SL run using all labeled data, a great amount of effort was saved by labeling only 100 sentences rather than 9,500. Co-training produced the best SSL results. For example, with only 200 labeled sentences, co-training yielded accuracy as high as 93.8%. Overall, SSL for opinion detection on movie reviews shows similar trends to SSL for traditional topical classification (Nigam and Ghani, 2000).

However, the advantages of SSL were not as significant in other data domains. Figure 1 demonstrates the performance of four types of SSL runs relative to corresponding baseline and full SL runs for all three data sets. All SSL runs reported here used 5% data as labeled data. Lines with different patterns indicate different data sets, green triangles mark baseline SL runs, green dots mark full SL runs, and red crosses mark SSL runs. Numbers next to symbols indicate classification accuracy. For each line, if the red cross is located above the triangle, it indicates that the SSL run improved over the SL baseline; and, the closer the red cross to the upper dot, the more effective was the SSL run. Figure 1 shows that S^{3}VM degrades in performance for all three data sets and we exclude it from the following discussion. From movie reviews to news articles to blog posts, the classification accuracy of baseline SL runs as well as the improvement gained by SSL runs decreased: With greater than 80% baseline accuracy on movie reviews, SSL runs were most effective; with slightly above 70% baseline accuracy on news articles, self-training actually decreased performance of the corresponding SL baseline while co-training and EM-NB outperformed the SL baseline only slightly; and with 60% or so baseline accuracy on blog posts, none of the SSL methods showed improvement. We assume that the lower the baseline accuracy, the worse the quality of auto-labeled data, and, therefore, the less advantages is application of SSL. We also found that the average sentence length in blog posts (17 words) is shorter than the average sentence length in either movie reviews (23.5 words) or news articles (22.5 words), which posed an additional challenge because there is less information for the classifier in terms of numbers of features.

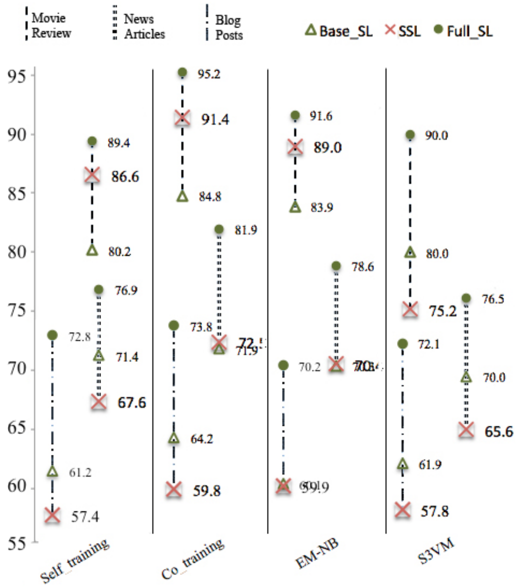

Figure 1: Classification accuracy(%) of SSL and SL on three data sets (i=5)

Overall, for movie reviews and news articles, co-training proved to be most robust and effective and EM-NB showed consistent improvement over the SL baseline. For news articles, EM-NB increased accuracy from 63.5% to 68.8% with only 100 labeled sentences. For movie reviews, a close look at EM-NB iterations shows that, with only 32 labeled sentences, EM-NB was able to achieve 88% classification accuracy, which is close to the best performance of simple Naïve Bayes self-training using 300 labeled sentences. This implies that the prob-

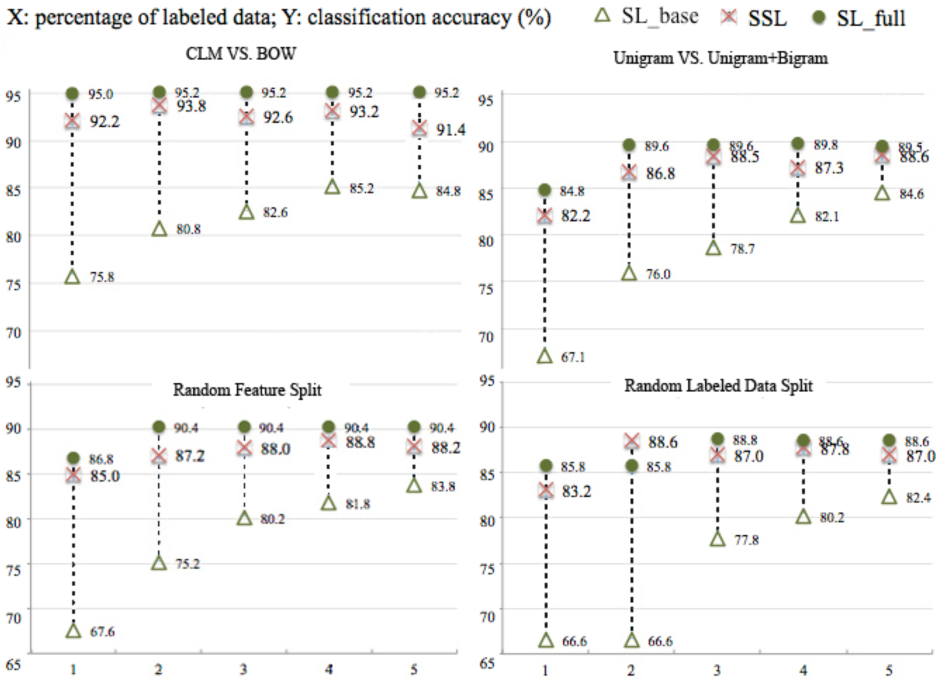

Figure 2: Performance of four co-training strategies on movie review data

lem space of opinion detection may be successfully described by the mixture model assumption of EM. As for blog posts, since the performance of the baseline classifiers was only slightly better than chance (50%), we needed to improve the baseline accuracy in order for SSL to work. One solution was to introduce high quality features. We augmented feature set with domain independent opinion lexicons that have been suggested as effective in creating high precision opinion classifier, but improvement was only minimal. An alternative solution was to borrow more labeled data from non-blog domains(s). Section 5.3 discusses dealing with a 'difficult' data domain using data from an 'easy' domain.

The preliminary exploration of different parameter settings for both self- and co-training showed no significant benefit gained by setting the weight parameter λ or applying feature selection; and using a larger number of unlabeled sentences u available for each iteration did not improve results. Further investigation is needed for an in-depth explanation. The poor performance of S^{3}VMs also needs to be investigated further.

5.2 Co-training

The best co-training runs reported in Table 1 and Figure 1 used a character-based language model (8-grams) to train one classifier and a bag-of-words model to train the other classifier. These two classifiers differ both in feature representation (i.e., character vs. word) and in learning algorithm (language model vs. pure statistical model). Figure 2 shows that for the movie review domain, other simple co-training configurations also produced promising results by using different feature sets (e.g., unigrams and the union of unigrams and bigrams, or randomly split feature sets) or different training sets. In the news domain, we observed similar trends. This shows the robustness and great potential of co-training. Because even with the logistic model to output probabilistic scores for the SVM classifier, the difference in probabilities was too small to select a small number of top predictions, adding an SVM classifier for co-training did not improve accuracy and is not discussed here.

An observation of the performance of self-training and co-training over iterations confirmed that co-training used labeled data more effectively for opinion detection than self-training, as suggested for traditional topical classification. We found that, overall, co-training produces better performance than self-training and reaches optimal performance faster. For instance, with 500 labeled sentences, a self-training run reached an optimal classification accuracy of 88.2% after adding 4,828 automatically annotated sentences for training, while the co-training run reached an optimal performance of 89.4% after adding only 2,588 sentences.

5.3 Domain Transfer

Even without any explicit domain adaptation methods, results indicate that simple self-training alone is promising for tackling domain transfer between the source domain movie reviews and the target domains news articles and blog posts.

Target domain news articles We used 9,500 labeled movie review sentences to train a Naïve Bayes classifier for news articles. Although this classifier produced a fairly good classification accuracy of 89.2% on movie review data, its accuracy was poor (64.1%) on news data (i.e., domain-transfer SL run), demonstrating the severity of the domain transfer problem. Self-training with Naïve Bayes using unlabeled data from the news domain (i.e., domain-transfer SSL run) improved the situation somewhat: it achieved a classification accuracy of 75.1% surpassing the domain-transfer SL run by more than 17%. To further understand how well SSL handles the domain transfer problem, a full in-domain SL run that used all labeled news sentences was also performed. This full SL run achieved 76.9% classification accuracy, only 1.8% higher than the domain-transfer SSL run, which did not use any labeled news data.

Target domain blog posts Because blog data are more challenging than news data, we kept 5% blog data as labeled data. Both SSL runs with and without out-of-domain data are depicted in Figure 3. Self-training using only blog data decreases SL baseline performance (dashed black line). Keeping the same settings, we added additional labeled data from the movie reviews, and self-training (gray line) came closer to the performance of the full SL run (red line), which used 90% of the labeled blog data. We then added a control factor that reduced the impact of movie review data gradually (i.e., a decrease of 0.001 in each iteration). Using this control, the self-training run (solid black line) reached and occasionally exceeded the performance of the full SL run.

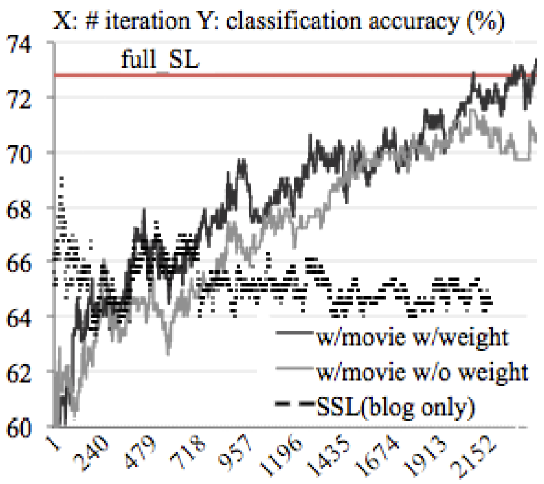

Figure 3: Self-training for domain transfer between movie reviews (source domain) and blogs (target domain)

6 Conclusion and Future Work

We investigated major SSL methods for identifying opinionated sentences in three domains. For movie review data, SSL methods attained state-of-the-art results with a small number of labeled sentences. Even without a natural feature split, different co-training strategies increased the baseline SL performance and outperformed other SSL methods. Due to the nature of the movie review data, we suspect that opinion detection on movie reviews is an 'easy' problem because it relies, strictly speaking, on distinguishing movie reviews from plot summaries, which also involves genre classification. For other manually created data sets that are expected to reflect real opinion characteristics, the SSL approach was impeded by low baseline precision and showed limited improvement. With the addition of out-of-domain labeled data, however, self-training exceeded full SL. This constitutes a successful new approach to domain adaptation.

Future work will include integrating opinion lexicons to bootstrap baseline precision and exploring co-training for domain adaptation.

References

Alias-i. 2008. LingPipe (version 4.0.1). Available from `http://alias-i.com/lingpipe`.

Anthony Aue and Michel Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 440–447, Prague, Czech Republic.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K. Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Menlo Park, CA.

Hagen Fürstenau and Mirella Lapata. 2009. Semi-supervised semantic role labeling. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 220–228, Athens, Greece.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Wei Jin, Hung Hay Ho, and Rohini K. Srihari. 2009. OpinionMiner: A novel machine learning system for web opinion mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The ICWSM 2010 JDPA sentiment corpus for the automotive domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC)*, Washington, D.C.

Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, McLean, VA.

Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. 1999. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sapporo, Japan.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2006. Latent variable models for semantic orientations of phrases. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy.

Partha Pratim Talukdar and Fernando Pereira. 2010. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1473–1481, Uppsala, Sweden.

Songbo Tan, Xueqi Cheng, Yufen Wang, and Hongbo Xu. 2009. Adapting naive Bayes to domain adaptation for sentiment analysis. In *Proceedings of the 31st European Conference on Information Retrieval (ECIR)*, Toulouse, France.

Wei Wang and Zhi-Hua Zhou. 2007. Analyzing co-training style algorithms. In *Proceedings of the 18th European Conference on Machine Learning*, Warsaw, Poland.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Mexico City, Mexico.

Janyce Wiebe, Rebecca Bruce, and Thomas O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, College Park, MD.

Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. 2001. A corpus study of evaluative and speculative language. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Kiduk Yang, Ning Yu, and Hui Zhang. 2007. WIDIT in TREC-2007 blog track: Combining lexicon-based methods to detect opinionated blogs. In *Proceedings of the 16th Text Retrieval Conference (TREC)*, Gaithersburg, MD.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing (EMNLP), Sapporo, Japan.

Wei Zhang and Clement Yu. 2007. UIC at TREC 2007 blog track. In *Proceedings of the 16th Text Retrieval Conference (TREC)*, Gaithersburg, MD.