Cross-Domain Co-Extraction of Sentiment and Topic Lexicons

Fangtao Li[§]**, Sinno Jialin Pan**[†]**, Ou Jin**[‡]**, Qiang Yang**[‡] **and Xiaoyan Zhu**[§]

[§]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[§]{fangtao06@gmail.com, zxy-dcs@tsinghua.edu.cn}
[†]Institute for Infocomm Research, Singapore
[†]jspan@i2r.a-star.edu.sg
[‡]Hong Kong University of Science and Technology, Hong Kong, China
[‡]{kingomiga@gmail.com, qyang@cse.ust.hk}

Abstract

Extracting sentiment and topic lexicons is important for opinion mining. Previous works have showed that supervised learning methods are superior for this task. However, the performance of supervised methods highly relies on manually labeled training data. In this paper, we propose a domain adaptation framework for sentiment- and topic- lexicon co-extraction in a domain of interest where we do not require any labeled data, but have lots of labeled data in another related domain. The framework is twofold. In the first step, we generate a few high-confidence sentiment and topic seeds in the target domain. In the second step, we propose a novel Relational Adaptive bootstraPping (RAP) algorithm to expand the seeds in the target domain by exploiting the labeled source domain data and the relationships between topic and sentiment words. Experimental results show that our domain adaptation framework can extract precise lexicons in the target domain without any annotation.

1 Introduction

In the past few years, opinion mining and sentiment analysis have attracted much attention in Natural Language Processing (NLP) and Information Retrieval (IR) (Pang and Lee, 2008; Liu, 2010). Sentiment lexicon construction and topic lexicon extraction are two fundamental subtasks for opinion mining (Qiu et al., 2009). A sentiment lexicon is a list of sentiment expressions, which are used to indicate sentiment polarity (e.g., positive or negative). The sentiment lexicon is domain dependent as users may use different sentiment words to express their opinion in different domains (e.g., different products). A topic lexicon is a list of topic expressions, on which

the sentiment words are expressed. Extracting the topic lexicon from a specific domain is important because users not only care about the overall sentiment polarity of a review but also care about which aspects are mentioned in review. Note that, similar to sentiment lexicons, different domains may have very different topic lexicons.

Recently, Jin and Ho (2009) and Li *et al.* (2010a) showed that supervised learning methods can achieve state-of-the-art results for lexicon extraction. However, the performance of these methods highly relies on manually annotated training data. In most cases, the labeling work may be time-consuming and expensive. It is impossible to annotate each domain of interest to build precise domain-dependent lexicons. It is more desirable to automatically construct precise lexicons in domains of interest by transferring knowledge from other domains.

In this paper, we focus on the co-extraction task of sentiment and topic lexicons in a target domain where we do not have any labeled data, but have plenty of labeled data in a source domain. Our goal is to leverage the knowledge extracted from the source domain to help lexicon co-extraction in the target domain. To address this problem, we propose a two-stage domain adaptation method. In the first step, we build a bridge between the source and target domains by identifying some *common* sentiment words as sentiment seeds in the target domain, such as "good", "bad", "nice", etc. After that, we generate topic seeds in the target domain by mining some *general* syntactic relation patterns between the sentiment and topic words from the source domain. In the second step, we propose a Relational Adaptive bootstraPping (RAP) algorithm to expand the seeds in the target domain. Our proposed method can uti-

lize useful labeled data from the source domain as well as exploit the relationships between the topic and sentiment words to propagate information for lexicon construction in the target domain. Experimental results show that our proposed method is effective for cross-domain lexicon co-extraction.

In summary, we have three main contributions: 1) We give a systematic study on cross-domain sentiment analysis in word level. While, most of previous work focused on document level; 2) A new two-step domain adaptation framework, with a novel RAP algorithm for seed expansion, is proposed. 3) We conduct extensive evaluation, and the experimental results demonstrate the effectiveness of our methods.

2 Related Work

2.1 Sentiment or Topic Lexicon Extraction

Sentiment or topic lexicon extraction is to identify the sentiment or topic words from text. In the past, many machine learning techniques have been proposed for this task. Hu and Liu *et al.* (2004) proposed an association-rule-based method to extract topic words and a dictionary-based method to identify sentiment words, independently. Wiebe *et al.* (2004) and Rioff *et al.* (2003) proposed to identify subjective adjectives and nouns using word clustering based on their distributional similarity. Popescu and Etzioni (2005) proposed a relaxed labeling approach to utilize linguistic rules for opinion polarity detection. Some researchers also proposed to use topic modeling to identify implicit topics and sentiment words (Mei et al., 2007; Titov and McDonald, 2008; Zhao et al., 2010; Li et al., 2010b), where a topic is a cluster of words, which is different from our fine-grained topic-word extraction.

Jin and Ho (2009) and Li *et al.* (2010a) both proposed to use supervised sequential labeling methods for topic and opinion extraction. Experimental results showed that the supervised learning methods can achieve state-of-the-art performance on lexicon extraction. However, these methods need to manually annotate a lot of training data in each domain. Recently, Qiu *et al.* (2009) proposed a rule-based semi-supervised learning methods for lexicon extraction. However, their method requires to manually define some *general* syntactic rules among sentiment and topic words. In addition, it still requires some annotated words in the target domain. In this paper, we do not assume any predefined rules and

labeled data be available in the target domain.

2.2 Domain Adaptation

Domain adaptation aims at transferring knowledge across domains where data distributions may be different (Pan and Yang, 2010). In the past few years, domain adaptation techniques have been widely applied to various NLP tasks, such as part-of-speech tagging (Ando and Zhang, 2005; Jiang and Zhai, 2007; Daumé III, 2007), named-entity recognition and shallow parsing (Daumé III, 2007; Jiang and Zhai, 2007; Wu et al., 2009). There are also lots of studies for cross-domain sentiment analysis (Blitzer et al., 2007; Tan et al., 2007; Li et al., 2009; Pan et al., 2010; Bollegala et al., 2011; He et al., 2011; Glorot et al., 2011). However, most of them focused on coarse-grained document-level sentiment classification, which is different from our fine-grained word-level extraction. Our work is similar to Jakob and Gurevych (2010) which proposed a Conditional Random Field (CRF) for cross-domain topic word extraction. However, the performance of their method highly depends on the manually designed features. In our experiments, we compare our method with theirs, and find that ours can achieve much better results on cross-domain lexicon extraction. Note that our work is also different from a recent work (Du et al., 2010), which focused on identifying the polarity of adjective words by using cross-domain knowledge. While we extract both topic and sentiment words and allow non-adjective sentiment words, which is more practical.

3 Cross-Domain Lexicon Co-Extraction

3.1 Problem Definition

Recall that, we focus on the setting where we have no labeled data in the target domain, while we have plenty of labeled data in the source domain. Denote $\mathcal{D}_S = \{(w_{S_i}, y_{S_i})\}_{i=1}^{n_1}$ the source domain data, where w_{S_i} represents a word in the source domain. $y_{S_i} \in \mathcal{Y}$ is the corresponding label of w_{S_i}. Similarly, we denote $\mathcal{D}_T = \{w_{T_j}\}_{j=1}^{n_2}$ the target domain data, where the input w_{T_j} is a word in the target domain. In lexicon extraction, $\mathcal{Y} \in \{1, 2, 3\}$, where $y_i = 1$ denotes the corresponding word w_i a sentiment word, $y_i = 2$ denotes w_i a topic word, and $y_i = 3$ denotes w_i neither a sentiment nor topic word. Our goal is to predict labels on $\mathcal{D}_T$ to extract topic and sentiment words for constructing topic and

sentiment lexicons, respectively.

3.2 Motivating Examples

In this section, we use some examples to introduce the motivation behind our proposed method. Table 1 shows several reviews from two domains: *movie* and *camera*. From the table, we can observe that there are some common sentiment words across different domains, such as "great", "excellent" and "amazing". However, the topic words may be different. For example, in the movie domain, topic words include "movie" and "script". While in the camera domain, topic words include "camera" and "photos".

Domain	Review
camera	The **camera** is *great*.
	it is a very *amazing* **product**.
	i highly *recommend* this **camera**.
	takes *excellent* **photos**.
	photos had some *artifacts* and *noise*.
movie	This **movie** has *good* **script**, *great* **casting**, *excellent* **acting**.
	I *love* this **movie**.
	Godfather was the most *amazing* **movie**.
	The **movie** is *excellent*.

Table 1: Reviews in *camera* and *movie* domains. Boldfaces are topic words and Italics are sentiment words.

Based on the observations, we can build a connection between the source and target domains by identifying the common sentiment words. Furthermore, intuitively, there are some general syntactic relationships or patterns between topic and sentiment words across different domains. Therefore, if we can mine the patterns from the source and target domain data, then we are able to construct an indirect connection between topic words across domains by using the common sentiment words as a bridge, which makes knowledge transfer across domains possible.

Figure 1 shows two dependency trees for the sentence "the camera is great" in the camera domain and the sentence "the movie is excellent" in the movie domain, respectively. As can be observed, the relationships between the topic and sentiment words in the two sentences are the same. They both share a "TOPIC-nsubj-SENTIMENT" relation. Let the camera domain be the source domain and the movie domain be the target domain. If the word "excellent" is identified as a common sentiment word, and the "TOPIC-nsubj-SENTIMENT" relation extracted from the camera domain is recognized as a common

syntactic pattern, then the word "movie" can be predicted as a topic word in the movie domain with high probability. After new topic words are extracted in the movie domain, we can apply the same syntactic pattern or other syntactic patterns to extract new sentiment and topic words iteratively.

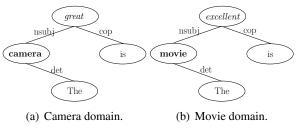
(a) Camera domain. (b) Movie domain.

Figure 1: Examples of dependency tree structure.

More specifically, we use the shortest path between a topic word and a sentiment word in the corresponding dependency tree to denote the relation between them. To get more general paths, we do not take original words in the path into consideration, but use their POS tags instead, such as "NN", "VB", "JJ", etc. As an example shown in Figure 2, we can extract two paths or relationships between topic and sentiment words from the dependency tree of the sentence "The movie has good script": "NN-amod-JJ" from "script" and "good", and "NN-nsubj-VB-dobj-NN-amod-JJ" from "movie" and "good".
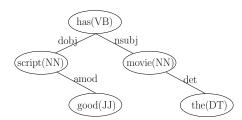

Figure 2: Example of pattern extraction.

In the following sections, we present the proposed two-stage domain adaptation framework: 1) generating some sentiment and topic seeds in the target domain; and 2) expanding the seeds in the target domain to construct sentiment and topic lexicons.

4 Seed Generation

Our basic idea is to first identify several *common* sentiment words across domains as sentiment seeds. Meanwhile, we mine some general patterns between sentiment and topic words from the source domain. Finally, we use the sentiment seeds and general patterns to generate topic seeds in the target domain.

4.1 Sentiment Seed Generation

To identify *common* sentiment words across domains, we extract all sentiment words from the source domain as candidates. For each candidate, we calculate its score based on the following metric:

$$S_1(w_i) = (p_S(w_i) + p_T(w_i)) \, e^{(-|p_S(w_i) - p_T(w_i)|)}, \quad (1)$$

where $p_S(w_i)$ and $p_T(w_i)$ are the probabilities of the word w_i occurring in the source and target domains, respectively. If a word w_i has high S_1 score, which implies that the word w_i occurs frequently and similarly in both domains, then it can be considered as a *common* sentiment word (Pan et al., 2010; Blitzer et al., 2007). We select top r candidates with highest S_1 scores as sentiment seeds.

4.2 Topic Seed Generation

We extract all patterns between sentiment and topic words in the source domain as candidates. For each pattern candidate, we calculate its score based on a metric defined in AutoSlog-TS (Riloff, 1996):

$$S_2(R_j) = Acc(R_j) \times log_2(Freq(R_j)), \quad (2)$$

where $Acc(R_j)$ is the accuracy of the pattern R_j in the source domain, and $Freq(R_j)$ is the frequency of the pattern R_j observed in target domain. This metric aims to identify the patterns that are precise in the source domain and observed frequently in the target domain. We also select the top r patterns with highest S_2 scores. With the patterns and sentiment seeds, we extract topic-word candidates and measure their scores based on a variant metric of quadratic combination (Zhang and Ye, 2008):

$$S_3(w_k) = \sum_{R_j \in A, \, w_i \in B} (S_2(R_j) \times S_1(w_i)), \quad (3)$$

where B is a set of sentiment seeds and A is a set of patterns which the words w_i and w_k satisfy. We then select the top r candidates as topic seeds.

5 Seed Expansion

After generating the topic and sentiment seeds, we aim to expand them in the target domain to construct topic and sentiment lexicons. In this section, we propose a new bootstrapping-based method to address this problem.

Bootstrapping is the process of improving the performance of a weak classifier by iteratively adding training data and retraining the classifier. More specifically, bootstrapping starts with a small set of labeled "seeds", and iteratively adds unlabeled data that are labeled by the classifier to the training set based on some selection criterion, and retrain the classifier. Many bootstrapping-based algorithms have been proposed to information extraction and other NLP tasks (Blum and Mitchell, 1998; Riloff and Jones, 1999; Jones et al., 1999; Wu et al., 2009).

One important issue in bootstrapping is how to design a criterion to select unlabeled data to be added to the training set iteratively. Our proposed bootstrapping for cross-domain lexicon extraction is based on the following two observations: 1) Although the source and target domains are different, part of source domain labeled data is still useful for lexicon extraction in the target domain after some adaptation; 2) The syntactic relationships among sentiment and topic words can be used to expand the seeds in the target domain for lexicon construction.

Based on the two observations, we propose a new bootstrapping-based method named Relational Adaptive bootstraPping (RAP), as summarized in Algorithm 1, for expanding lexicons across domains. In each iteration, we employ a cross-domain classifier trained on the source domain lexicons and the extracted target domain lexicons to predict the labels of the target unlabeled data, and select top k_2 predicted topic and sentiment words as candidates based on confidence. With the extracted syntactic patterns in the previous iterations, we construct a bipartite graph between sentiment and topic words on the extracted target domain lexicons and candidates. After that, a graph-based score refinement algorithm is performed on the graph, and the top k_1 candidates are added to the extracted lexicons based on the final scores. Accordingly, with the new extracted lexicons, we update the syntactic patterns in each iteration. The details of RAP are presented in the following sections.

5.1 Cross-Domain Classifier

In this paper, we employ *Transfer AdaBoost* (TrAdaBoost) (Dai et al., 2007) as the cross-domain learning algorithm in RAP. In TrAdaBoost, each word w_{S_i} (or w_{T_j}) is represented by a feature vector x_{S_i} (or x_{T_j}). A classifier trained on the source domain data $\mathcal{D}_S = \{(x_{S_i}, y_{S_i})\}$ may perform poor on x_{T_j} because of domain difference. The main idea of TrAdaBoost is to re-weight the source domain data based on a few of target domain labeled data, which is referred to as seeds in our task. The re-weighting

aims to reduce the effect of the "bad" source domain data while encourage the "good" ones to get a more precise classifier in target domain. In each iteration of RAP, we train cross-domain classifiers $f_{\mathcal{O}}^T$ and $f_{\mathcal{P}}^T$ for sentiment- and topic- word extraction using TrAdaBoost separately (taking sentiment or topic words as positive instances). We use linear Support Vector Machines (SVMs) as the base classifier in TrAdaBoost. For features to represent each word, we use lexicon features, such as the previous, current and next words, and POS tag features, such as the previous, current and next words' POS tags.

Algorithm 1 Relational Adaptive bootstraPping

Require: Target domain data $\mathcal{D}_T = \mathcal{D}_T^l \bigcup \mathcal{D}_T^u$, where $\mathcal{D}_T^l$ consists of sentiment seeds B and topic seeds C and their initial scores $S_1(w_i), \forall w_i \in B$ and $S_3(w_j), \forall w_j \in C, \mathcal{D}_T^u$ is the set of unlabeled target domain data; labeled source domain data $\mathcal{D}_S$; a cross-domain classifier; iteration number M and candidate selection number k_1, k_2.

Ensure: Expand C and B in the target domain.

1: Initialize a pattern set $A = \emptyset$, $\widetilde{S}_1(w_i) = S_1(w_i), w_i \in B$ and $\widetilde{S}_3(w_j) = S_3(w_j), w_j \in C$. Consider all patterns observed in the source domain as pattern candidates P.
2: **for** $m = 1 \ldots M$ **do**
3:   Extract new pattern candidates to P with $\mathcal{D}_T^l$ in target domain, update pattern score $\widetilde{S}_2(R_j)$, where $R_j \in P$, based on Eq. (4), and select the top k_1 patterns to the pattern set A.
4:   Learn the cross-domain classifiers $f_{\mathcal{O}}^T$ and $f_{\mathcal{P}}^T$ for sentiment- and topic- word extraction with $\mathcal{D}_S \bigcup \mathcal{D}_T^l$ separately. Predict the sentiment score $h_{f_{\mathcal{O}}}^T(w_{T_j})$ and topic score $h_{f_{\mathcal{P}}}^T(w_{T_j})$ on $\mathcal{D}_T^u$, and select k_2 sentiment words and topic words with highest scores as candidates.
5:   Construct a bipartite graph between sentiment and topic words on $\mathcal{D}_T^l$ and the k_2 sentiment- and topic- word candidates, and calculate the normalized weights θ_{ij}'s for each edge of the graph.
6:   Refine the scores $\widetilde{S}_1$ and $\widetilde{S}_3$ of the k_2 sentiment and topic word candidates using Eqs. (5) and (6) iteratively.
7:   Select k_1 new sentiment words and k_1 new topic words with the final scores, and add them to lexicons B and C. Update $\widetilde{S}_1(w_i)$ and $\widetilde{S}_3(w_j)$ accordingly.
8: **end for**
9: **return** Expanded lexicons B and C.

5.2 Graph Construction

Based on the cross-domain classifiers $f_{\mathcal{O}}^T$ and $f_{\mathcal{P}}^T$, we can predict the sentiment label score $h_{f_{\mathcal{O}}}^T(w_{T_i})$ and topic label score $h_{f_{\mathcal{P}}}^T(w_{T_i})$ for the target domain data w_{T_i}. According to all predicted values, we respectively select top k_2 new sentiment- and topic- words as candidates. Together with the extracted sentiment and topic lexicons in the target domain,

we build a bipartite graph among them as shown in Figure 3. In the bipartite graph, one set of nodes represents topic words, including new topic candidates and words in the lexicon C, and the other set of nodes represents sentiment words, including new sentiment candidates and words in the lexicon B. For a pair of sentiment and topic words $w_{T_i}^{\mathcal{O}}$ and $w_{T_j}^{\mathcal{P}}$, if there is a pattern R_j in the pattern set A that they can satisfy, then there exists an edge e_{ij} between them. Furthermore, each edge e_{ij} is associated with a nonnegative weight θ_{ij}, which is measured as follows, $\theta_{ij} = \sum_{R_k \in E} \widetilde{S}_2(R_k)$, where $\widetilde{S}_2$ is the pattern score. Similar to the metric defined in Eq. (3), the pattern score is defined as:

$$\widetilde{S}_2(R_j) = \sum_{\{w_i, w_k\} \in E} \left(\widetilde{S}_1(w_i) \times \widetilde{S}_3(w_k) \right), \quad (4)$$

where $E = \{\{w_i, w_j\}|, w_i \in B, w_j \in C$ and w_i, w_j satisfy $R_j, R_j \in A\}$. Note that in the beginning of each iteration, $\widetilde{S}_2$ is updated based on the new sentiment score $\widetilde{S}_1$ and topic score $\widetilde{S}_3$. We further normalize θ_{ij} by $\widetilde{\theta}_{ij} = \theta_{ij}/(\sum_{ij} \theta_{ij})$.
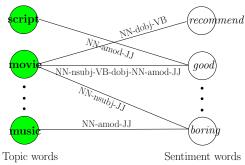

Figure 3: Topic and sentiment word graph.

5.3 Score Computation

We construct the bipartite graph to exploit the relationships between sentiment and topic words to propagate information for lexicon extraction. We use the following reinforcement formulas to iteratively update the final sentiment score $\widetilde{S}_1(w_{T_j})$ and topic score $\widetilde{S}_3(w_{T_i})$, respectively:

$$\widetilde{S}_1(w_{T_j}) = \mu \sum_i \widetilde{S}_3(w_{T_i}) \widetilde{\theta}_{ij} + (1 - \mu) h_{f_{\mathcal{O}}}^T(w_{T_j}), \quad (5)$$

$$\widetilde{S}_3(w_{T_i}) = \mu \sum_j \widetilde{S}_1(w_{T_j}) \widetilde{\theta}_{ij} + (1 - \mu) h_{f_{\mathcal{P}}}^T(w_{T_i}), \quad (6)$$

where μ is a trade-off parameter between the predicted value by cross-domain classifier and the reinforcement scores from other nodes connected by

edge e_{ij}. Here μ is empirically set to be 0.5. With Eqs. (5) and (6), the sentiment scores and topic scores are iteratively refined until the state of the graph trends to be stable. This can be considered as an extension to the HITS algorithm(Kleinberg, 1999). Finally, we select $k_1 \ll k_2$ sentiment and topic words from the k_2 candidates based on their refined scores, and add them to the target domain lexicons, respectively. We also update the sentiment score $\widetilde{S}_1$ and topic score $\widetilde{S}_3$ for next iteration.

5.4 Special Cases

We now introduce two special cases of the RAP algorithm. In Eqs. (5) and (6), if the parameter $\mu = 1$, then RAP only uses the relationships between sentiment and topic words with their patterns to propagate label information in the target domain without using the cross-domain classifier. We call this reduction relational bootstrapping. If $\mu = 0$, then RAP only utilizes useful source domain labeled data to assist learning of the target domain classifier without considering the relationships between sentiment and topic words. We call this reduction adaptive bootstrapping, which can be considered as a bootstrapping version of TrAdaBoost. We also empirically study these two special cases in experiments.

6 Experiments on Lexicon Evaluation

6.1 Data Set and Evaluation Criteria

We use the review dataset from (Li et al., 2010a), which contains 500 movie and 601 product reviews, for evaluation. The sentiment and topic words are manually annotated. In this dataset, all types of sentiment words are annotated instead of adjective words only. For example, the verbs, such as "like", "recommend", and nouns, such as "masterpiece", are also labeled as sentiment words. We construct two cross-domain lexicon extraction tasks: "product vs. movie" and "movie vs. product", where the word before "vs." corresponds with the source domain and the word after "vs." corresponds with the target domain. We evaluate our methods in terms of precision, recall and F-score ($F1$).

6.2 Baselines

The results of in-domain classifiers, which are trained on plenty of target domain labeled data, can be treated as upper-bounds. We denote iSVM and iCRF the in-domain SVM and CRF classifiers in experiments, and compare our proposed methods,

RAP, relational bootstrapping, and adaptive bootstrapping, with the following baselines,

Unsupervised Method (Un) we implement a rule-based method for lexicon extraction based on (Hu and Liu, 2004), where adjective words that match a rule is recognized as sentiment words, and nouns that match a rule are recognized as topic words.

Semi-Supervised Method (Semi) we implement the double propagation model proposed in (Qiu et al., 2009). Since this method requires some target domain labeled data, we manually label 30 sentiment words in the target domain.

Cross-Domain CRF (Cross-CRF) we implement a cross-domain CRF algorithm proposed by (Jakob and Gurevych, 2010).

TrAdaBoost We apply TrAdaBoost (Dai et al., 2007) on the source domain labeled data and the generated seeds in the target domain to train a lexicon extractor.

6.3 Comparison Results

Comparison results on lexicon extraction are shown in Table 2 and Table 3. From Table 2, we can observe that our proposed methods are effective for sentiment lexicon extraction. The relational bootstrapping method performs better than the unsupervised method, TrAdaBoost and the cross-domain CRF algorithm, and achieves comparable results with the semi-supervised method. However, compared to the semi-supervised method, our proposed relational bootstrapping method does not require any labeled data in the target domain. We can also observe that the adaptive bootstrapping method and the RAP method perform much better than other methods in terms of F-score. The reason is that part of the source domain labeled data may be useful for learning the target classifier after reweighting. In addition, we also observe that embedding the TrAdaBoost algorithm into a bootstrapping process can further boost the performance of the classifier for sentiment lexicon extraction.

Table 3 shows the comparison results on topic lexicon extraction. From the table, we can observe that different from the sentiment lexicon extraction task, the relational bootstrapping method performs better than the adaptive bootstrapping method slightly. The reason may be that for the sentiment lexicon extraction task, there exist some common sentiment words

	product vs. movie			movie vs. product		
	Prec.	Rec.	$F1$	Prec.	Rec.	$F1$
Un	0.82	0.31	0.45	0.74	0.23	0.35
Semi	0.71	0.44	0.54	0.62	0.45	0.52
Cross-CRF	0.69	0.40	0.51	0.65	0.34	0.45
Tradaboost	0.73	0.41	0.52	0.72	0.42	0.52
Adaptive	0.68	0.53	0.59	0.63	0.52	0.57
Relational	0.55	0.51	0.53	0.57	0.51	0.54
RAP	0.69	0.59	**0.64**	0.66	0.59	**0.62**
iSVM	0.82	0.60	0.70	0.80	0.61	0.68
iCRF	0.80	0.66	0.72	0.80	0.62	0.69

Table 2: Results on sentiment lexicon extraction. Numbers in boldface denote significant improvement.

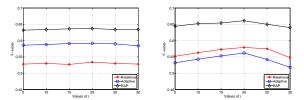|	product vs. movie			movie vs. product		
	Prec.	Rec.	$F1$	Prec.	Rec.	$F1$
Un	0.41	0.32	0.36	0.53	0.35	0.41
Semi	0.54	0.59	0.56	0.75	0.50	0.60
Cross-CRF	0.70	0.23	0.34	0.80	0.24	0.37
Tradaboost	0.64	0.45	0.53	0.57	0.47	0.51
Adaptive	0.76	0.44	0.56	0.70	0.52	0.59
Relational	0.57	0.58	0.58	0.61	0.57	0.59
RAP	0.80	0.56	**0.66**	0.73	0.58	**0.65**
iSVM	0.83	0.73	0.78	0.85	0.70	0.77
iCRF	0.84	0.78	0.81	0.87	0.73	0.80

Table 3: Results on topic lexicon extraction. Numbers in boldface denote significant improvement.

across domains, thus part of the labeled source domain data may be useful for the target learning task. However, for the topic lexicon extraction task, the topic words may be totally different, and as a result, we may not be able to find useful source domain labeled data to boost the performance for lexicon extraction in the target domain. In this case, mutual label propagation between sentiment and topic words may be more reasonable for knowledge transfer. RAP absorbs the advantages of the adaptive and relational bootstrapping methods, thus can get the best results in both lexicon extraction tasks.
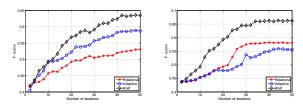
We also observe that relational bootstrapping can get better recall, but lower precision, compared to adaptive bootstrapping. This is because relational bootstrapping only utilizes the patterns to propagate label information, which may cover more topic and sentiment seeds, but include some *noisy* words. For example, given two phases "like the camera" and "recommend the camera", we can extract a pattern "VB-dobj-NN". However, by using this pattern and the topic word "camera", we may extract "take" as a sentiment word from another phase "take the cam-

era", which is incorrect. The adaptive bootstrapping method can utilize various features to make predictions more precisely, which may have higher precision, but encounter the lower recall problem. For example, "flash" is not identified as a topic word in the target product domain (camera domain). Our RAP method can exploit both relationships between sentiment and topic words and part of labeled source domain data for cross-domain lexicon extraction. It can correctly identify the above two cases.

6.3.1 Parameter Sensitivity Study

In this section, we conduct experiments to study the effect of different parameter settings. There are several parameters in the framework: the number of generated seeds r, the number of new candidates k_2 and the number of selections k in each iteration, and the number of iterations M (μ is empirically set to 0.5). For the parameter k_2, we just set it to a large number ($k_2 = 100$) such that have rich candidates to build the bipartite graph. In the experiments reported in the previous section, we set $r = 20$, $k_1 = 10$ and $M = 50$. Figures 4(a) and 4(b) show the results under varying values of r in the "product vs. movie" task. Observe that for sentiment word extraction, the results of the proposed methods are not sensitive to the values of r. While for the topic word extraction, the proposed methods perform well when r falls in the range from 15 to 20.

(a) Sentiment word extraction (b) Topic word extraction

Figure 4: Results on varying values of r.

(a) Sentiment word extraction (b) Topic word extraction

Figure 5: Results on varying values of M.

We also test the sensitivity of the parameter k_1 and find that the proposed methods work well and robust when k_1 falls in the range from 10 to 20.

Figures 5(a) and 5(b) show the results under varying numbers of iterations in the "product vs. movie" task. As we can see, our proposed methods converge well when $M \geq 40$.

7 Application: Sentiment Classification

To further verify the usefulness of the lexicons extracted by the RAP method, we apply the extracted sentiment lexicon for sentiment classification.

7.1 Experiment Setting

Our work is motivated by the work of (Pang and Lee, 2004), which only used subjective sentences for document-level sentiment classification, instead of using all sentences. In this experiment, we only use sentiment related words as features to represent opinion documents for classification, instead of using all words. Our goal is compare the sentiment lexicon constructed by the RAP method with other general lexicons on the impact of for sentiment classification. The general lexicons used for comparison are described in Table 4.

We use the dataset from (Blitzer et al., 2007) for sentiment classification. It contains a collection of product reviews from Amazon.com. The reviews are about four product domains: books, dvds, electronics and kitchen appliance. In each domain, there are 1000 positive and 1000 negative reviews. To construct domain specific sentiment lexicons, we apply RAP on each product domain with the *movie* domain described in Section 6.1 as the source domain. Finally, we use linear SVM as the classifier and the classification accuracy as the evaluate criterion.

Lexicon Name	Size	Description
Senti-WordNet	6957	Words with a subjective score > 0.6 (Esuli and Sebastiani, 2006)
HowNet	4619	Eng. translation of subj. Chinese words (Dong and Dong, 2006)
Subj. Clues	6878	Lexicons from (Wilson et al., 2005)

Table 4: Description of different lexicons.

7.2 Experimental Results

Experimental results on sentiment classification are shown in Table 5, where we denote "All" using all unigram and bigram features instead of using subjective words. As we can see that a classifier trained with features constructed by our RAP method performance best in all domains. Note that the number of features (sentiment words) constructed by our

method is much smaller than that of all unigram and bigram features, which can reduce the classifier training time dramatically. These promising results imply that our RAP can be applied for sentiment classification effectively and efficiently.

	All	Senti	HowNet	Subj. Clue	Ours
dvd	82.55	79.80	80.57	80.93	**84.05**
book	80.71	76.22	78.22	79.48	**81.65**
electronic	84.43	82.42	83.05	83.22	**86.71**
kitchen	87.70	81.78	84.17	84.23	**88.83**

Table 5: Sentiment classification results (accuracy in %). Numbers in boldface denotes significant improvement.

8 Conclusions

In this paper, we propose a two-stage framework for co-extraction of sentiment and topic lexicons across domains where we have no labeled data in the target domain but have plenty of labeled data in another domain. In the first stage, we propose a simple strategy to generate a few high-quality sentiment and topic seeds for the target domain. In the second stage, we propose a novel Relational Adaptive bootstraPping (RAP) method to expand the seeds, which can exploit the relationships between topic and opinion words, and make use of part of useful source domain labeled data for help. Extensive experimental results show our proposed method can extract precise sentiment and topic lexicons from the target domain. Furthermore, the extracted sentiment lexicon can be applied to sentiment classification effectively.

In the future work, besides the heterogeneous relationships between topic and sentiment words, we intend to investigate the homogeneous relationships among topic words and those among sentiment words (Qiu et al., 2009) to further boost the performance of RAP method. Furthermore, in our framework, we do not identify the polarity of the extracted sentiment lexicon. We also plan to embed this component into our unified framework. Finally, it is also interesting to exploit multi-domain knowledge (Li and Zong, 2008; Bollegala et al., 2011) for cross-domain lexicon extraction.

9 Acknowledgement

References

Rie K. Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853.

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*, pages 432–439, Prague, Czech Republic. ACL.

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.

Danushka Bollegala, David Weir, and John Carroll. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 132–141, Portland, Oregon. ACL.

Wenyuan Dai, Qiang Yang, Guirong Xue, and Yong Yu. 2007. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 193–200, Corvalis, Oregon, USA, June. ACM.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. ACL.

Zhendong Dong and Qiang Dong, editors. 2006. *HOWNET and the computation of meaning*. World Scientific Publishers, Norwell, MA, USA.

Weifu Du, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. 2010. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *Proceedings of the 3rd ACM international conference on Web search and data mining*, pages 111–120, New York, NY, USA. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTI-WORDNET: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, pages 513–520, Bellevue, Washington, USA.

Yulan He, Chenghua Lin, and Harith Alani. 2011. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 123–131, Portland, Oregon. ACL.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, Seattle, WA, USA. ACM.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045, Cambridge, Massachusetts, USA. ACL.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic. ACL.

Wei Jin and Hung Hay Ho. 2009. A novel lexicalized HMM-based learning framework for web opinion mining. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 465–472, Montreal, Quebec, Canada. ACM.

Rosie Jones, Andrew Mccallum, Kamal Nigam, and Ellen Riloff. 1999. Bootstrapping for text learning tasks. In *In IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, pages 52–63.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, Sept.

Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 257–260, Columbus, Ohio, USA. ACL.

Tao Li, Vikas Sindhwani, Chris Ding, and Yi Zhang. 2009. Knowledge transformation for cross-domain sentiment classification. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 716–717, Boston, MA, USA. ACM.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010a. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 653–661, Beijing, China.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010b. Sentiment analysis with global topics and local dependency. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, Atlanta, Georgia, USA. AAAI Press.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, Second Edition*.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180, Banff, Alberta, Canada. ACM.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, Oct.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Chen Zheng. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, pages 751–760, Raleigh, NC, USA, Apr. ACM.

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. ACL.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, British Columbia, Canada. ACL.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1199–1204, Pasadena, California, USA. Morgan Kaufmann Publishers Inc.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 6th national conference on Artificial intelligence*, pages 474–479, Orlando, Florida, United States. AAAI.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th conference on natural language learning*, pages 25–32, Edmonton, Canada. ACL.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1044–1049, Portland, Oregon, USA. AAAI Press/MIT Press.

Songbo Tan, Gaowei Wu, Huifeng Tang, and Xueqi Cheng. 2007. A novel scheme for domain-transfer problem in the context of sentiment analysis. In *Proceedings of the 16th ACM conference on Conference on information and knowledge management*, pages 979–982, Lisbon, Portugal. ACM.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*, pages 308–316, Columbus, Ohio, USA. ACL.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Comput. Linguist.*, 30:277–308, Sept.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. ACL.

Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. 2009. Domain adaptive bootstrapping for named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1523–1532, Singapore. ACL.

Min Zhang and Xingyao Ye. 2008. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 411–418, Singapore. ACM.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65, Cambridge, Massachusetts, USA. ACL.