Learning Latent Word Representations for Domain Adaptation using Supervised Word Clustering

Min Xiao and **Feipeng Zhao** and **Yuhong Guo**
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122, USA
{minxiao,feipeng.zhao,yuhong}@temple.edu

Abstract

Domain adaptation has been popularly studied on exploiting labeled information from a source domain to learn a prediction model in a target domain. In this paper, we develop a novel representation learning approach to address domain adaptation for text classification with automatically induced discriminative latent features, which are generalizable across domains while informative to the prediction task. Specifically, we propose a hierarchical multinomial Naive Bayes model with latent variables to conduct supervised word clustering on labeled documents from both source and target domains, and then use the produced cluster distribution of each word as its latent feature representation for domain adaptation. We train this latent graphical model using a simple expectation-maximization (EM) algorithm. We empirically evaluate the proposed method with both cross-domain document categorization tasks on Reuters-21578 dataset and cross-domain sentiment classification tasks on Amazon product review dataset. The experimental results demonstrate that our proposed approach achieves superior performance compared with alternative methods.

1 Introduction

Supervised prediction models typically require a large amount of labeled data for training. However, manually collecting data annotations is expensive in many real-world applications such as document categorization or sentiment classification. Recently, domain adaptation has been proposed to exploit existing labeled data in a related source domain to assist the prediction model training in the target domain (Ben-David et al., 2006; Blitzer et al., 2006; Daumé III, 2007; Blitzer et al., 2011; Chen et al., 2012). As an effective tool to reduce annotation effort, domain adaptation has achieved success in various cross-domain natural language processing (NLP) systems such as document categorization (Dai et al., 2007), sentiment classification (Blitzer et al., 2007; Chen et al., 2012; Mejova and Srinivasan, 2012; Chen et al., 2011), email spam detection (Jiang and Zhai, 2007), and a number of other NLP tasks (Blitzer et al., 2011; Daumé III, 2007).

One primary challenge of domain adaptation lies in the distribution divergence of the two domains in the original feature representation space. For example, documents about *books* may contain very different high-frequency words and discriminative words from documents about *kitchen*. A good cross-domain feature *representation* thus has been viewed as critical for bridging the domain divergence gap and facilitating domain adaptation in the NLP area (Ben-David et al., 2006, 2010). Many domain adaptation works have been proposed to learn new cross-domain feature representations (Blitzer et al., 2006, 2011). Though demonstrated good performance on certain problems, these works mostly induce new feature representations in an unsupervised way, without taking the valuable label information into account.

In this work, we present a novel supervised representation learning approach to discover a latent representation of words which is not only generalizable across domains but also informative to the classification task. Specifically, we propose a hier-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 152–162,
Seattle, Washington, USA, 18-21 October 2013.

archical multinomial Naive Bayes model with latent word cluster variables to perform supervised word clustering on labeled documents from both domains. Our model directly models the relationships between the observed document label variables and the latent word cluster variables. The induced cluster representation of each word thus will be informative for the classification labels, and hence discriminative for the target classification task. We train this directed graphical model using an expectation-maximization (EM) algorithm, which maximizes the log-likelihood of the observations of labeled documents. The induced cluster distribution of each word can then be used as its generalizable representation to construct new cluster-based representation of each document. For domain adaptation, we train a supervised learning system with labeled data from both domains in the new representation space and apply it to categorize test documents in the target domain. In order to evaluate the proposed technique, we conduct extensive experiments on the Reuters-21578 dataset for cross-domain document categorization and on Amazon product review dataset for cross-domain sentiment classification. The experimental results show the proposed approach can produce more effective representations than the comparison domain adaptation methods.

2 Related Work

Domain adaptation has recently been popularly studied in natural language processing and a variety of domain adaptation approaches have been developed, including instance weighting adaptation methods and feature representation learning methods.

Instance weighting adaptation methods improve the transferability of a prediction model by training an instance weighted learning system. Much work in this category has been developed to address different weighting schemas (Sugiyama et al., 2007; Wan et al., 2011). Jiang and Zhai (2007) applied instance weighting algorithms to tackle cross-domain NLP tasks and proposed to remove misleading source training data and assign less weights to labeled data from the source domain than labeled data from the target domain. Dai et al. (2007) proposed to increase the weights of mistakenly predicted instances from the target domain and decrease the weights of incorrectly predicted instances from the source domain during an iterative training process.

Representation learning methods bridge domain divergence either by differentiating domain-invariant features from domain-specific features (Daumé III, 2007; Daumé III et al., 2010; Blitzer et al., 2011; Finkel and Manning, 2009) or seeking generalizable latent features across domains (Blitzer et al., 2006, 2007; Prettenhofer and Stein, 2010). Daumé III (2007); Daumé III et al. (2010) proposed a simple heuristic feature replication method to represent common, source specific and target specific features. Finkel and Manning (2009) proposed a former version of it based on the use of a hierarchical Bayesian prior. Blitzer et al. (2011) proposed a coupled subspace learning method, which learns two projectors, one for each domain, to project the original features into domain-sharing and domain-specific features. Blitzer et al. (2006) proposed a structural correspondence learning (SCL) method to model the correlation between pivot features and non-pivot features. It uses the correlation to induce latent domain-invariant features as augmenting features for supervised learning. Extensions of this work include improving pivot feature selection (Blitzer et al., 2007; Prettenhofer and Stein, 2010), and improving the correlation modeling between pivot and non-pivot features (Tan, 2009).

The proposed approach in this paper belongs to representation learning methods. However, unlike the unsupervised representation learning methods reviewed above, our proposed approach learns generalizable feature representations of words by exploiting data labels from the two domains.

3 Learning Latent Word Representations using Supervised Word Clustering

In this paper, we address domain adaptation for text classification. Given a source domain $\mathcal{D}_{\mathcal{S}}$ with plenty of labeled documents, and a target domain $\mathcal{D}_{\mathcal{T}}$ with a very few labeled documents, the task is to learn a classifier from the labeled documents in both domains, and use it to classify the unlabeled documents in the target domain. The documents in the two domains share the same universal vocabulary $\mathcal{V} = \{w_1, w_2, \cdots, w_n\}$, but the word distributions in the two domains are typically different.

Therefore, training the classification model directly from the original word feature space $\mathcal{V}$ may not generalize well in the target domain.

We propose to address this problem by first learning a supervised mapping function $\phi : \mathcal{V} \longrightarrow \mathcal{Z}$ from the labeled documents in both domains, which maps the input word features in the large vocabulary set $\mathcal{V}$ into a low dimensional latent feature space $\mathcal{Z}$. By filtering out unimportant details and noises, we expect the low dimensional mapping can capture the intrinsic structure of the input data that is discriminative for the classification task and generalizable across domains. In particular, we learn such a mapping function by conducting supervised word clustering on the labeled documents using a hierarchical multinomial Naive Bayes model. Below, we will first introduce this supervised word clustering model and then use the mapping function produced to transform documents in different domains into the same low-dimensional space for training cross domain text classification systems.

3.1 Supervised Word Clustering

Given all labeled documents from the source and target domains, $D = \{(\mathbf{w}_t, y_t)\}_{t=1}^{T}$, where the t-th labeled document is expressed as a bag of words, $\mathbf{w}_t = \{w_{t1}, w_{t2}, \cdots, w_{tN_t}\}$, and its label value is $y_t \in \mathcal{Y}$ for $\mathcal{Y} = \{1, \cdots, K\}$, we propose to perform supervised word clustering by modeling the document-label pair distribution using a hierarchical multinomial Naive Bayes model given in Figure 1, which has a middle layer of latent cluster variables.

In this plate model, the variable Y_t denotes the observed class label for the t-th document, and all the label variables, $\{Y_t\}_{t=1}^{T}$, share the same multinomial distribution θ_Y across documents. The latent variable $C_{t,i}$ denotes the cluster membership of the word $W_{t,i}$, and all the cluster variables, $\{C_{t,i}\}_{t=1,i=1}^{T,N_t}$, share the same set of conditional distributions $\{\theta_{C|y}\}_{y=1}^{K}$ across documents and words. The variable $W_{t,i}$ denotes the i-th observed word in the t-th document, and all the word variables, $\{W_{t,i}\}_{t=1,i=1}^{T,N_t}$, share the same set of conditional distributions $\{\theta_{W|c}\}_{c=1}^{m}$. Here we assume the number of word clusters is m. For simplicity, we do not show the distribution parameter variables in the Figure.

Following the *Markov property* of directed graphical models, we can see that given the cluster variable values, the document label variables will be completely independent of the word variables. By learning this latent directed graphical model, we thus expect the important classification information expressed in the input observation words can be effectively summarized into the latent cluster variables. This latent model is much simpler than the supervised topic models (Blei and McAuliffe, 2007), but we will show later that it can suitably produce a generalizable feature mapping function for domain adaptation.

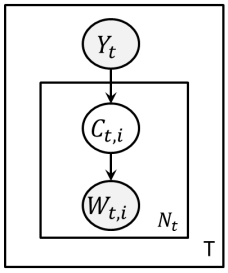

Figure 1: Supervised word clustering model.

To train the latent graphical model in Figure 1 on labeled documents D, we use a standard expectation-maximization (EM) algorithm (Dempster et al., 1977) to maximize the marginal log-likelihood of the observations:

$$LL(D; \boldsymbol{\theta}) = \sum_{t} \log P(y_t, \mathbf{w}_t | \boldsymbol{\theta}) \qquad (1)$$

The EM algorithm is an iterative procedure. In each iteration, it takes an alternative E-step and M-step to maximize the lower bound of the marginal log-likelihood function. In our experiments, we start from a random initialization of the model parameters and the latent variable values, and then perform iterative EM updates until converge to a local optimal solution.

3.2 Induced Word Representation

After training the supervised clustering model using EM algorithm, a set of local optimal model parameters $\boldsymbol{\theta}^*$ will be returned, which define a joint distribution over the three groups of variables in the directed graphical model. Next we define a supervised latent feature mapping function ϕ from this trained

model to map each word w in the vocabulary $\mathcal{V}$ into a conditional distribution vector over the word cluster variable, such as

$$\phi(w) = [P(c=1|w, \boldsymbol{\theta}^*), \cdots, P(c=m|w, \boldsymbol{\theta}^*)]. \quad (2)$$

The conditional distributions involved in this mapping function can be computed as

$$P(c|w, \boldsymbol{\theta}^*) = \frac{\sum_{y \in \mathcal{Y}} P(w|c, \boldsymbol{\theta}^*) P(c|y, \boldsymbol{\theta}^*) P(y|\boldsymbol{\theta}^*)}{P(w)} \quad (3)$$

where $P(w|c, \boldsymbol{\theta}^*) = \boldsymbol{\theta}^*_{w|c}$ $P(c|y, \boldsymbol{\theta}^*) = \boldsymbol{\theta}^*_{c|y}$ and $P(y|\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*_y$ can be determined from the model parameters directly, and $p(w)$ can be computed as the empirical frequency of word w among all the other words in all the training documents.

We then define a transformation matrix $\Pi \in \mathbb{R}^{n \times m}$ based on the mapping function ϕ defined in Eq. (2), such that $\Pi_{i:} = \phi(w_i)$ where w_i is the i-th word in the vocabulary $\mathcal{V}$. That is, each row of Π is the induced representation vector for one word. Π can be viewed as a soft word clustering matrix, and $\Pi_{i,j}$ denotes the probability of word w_i belongs to the j-th cluster. Given the original document-word frequency matrix $X_{tr} \in \mathbb{R}^{T \times n}$ for the labeled training documents from the two domains, we can construct its representations $Z_{tr} \in \mathbb{R}^{T \times m}$ in the predictive latent clustering space by performing the following transformation:

$$Z_{tr} = X_{tr} \Pi. \quad (4)$$

Similarly, we can construct the new representation matrix Z_{ts} for the test data X_{ts} in the target domain. We then train a classification model on the labeled data Z_{tr} and apply it to classify the test data Z_{ts}.

4 Experiments

We evaluate the proposed approach with experiments on cross domain document categorization of Reuters data and cross domain sentiment classification of Amazon product reviews, comparing to a number of baseline and existing domain adaptation methods. In this section, we report the experimental setting and results on these two data sets.

4.1 Approaches

We compared our proposed supervised word clustering approach (*SWC*) with the following five comparison methods for domain adaptation:

(1) *BOW*: This is a bag-of-word baseline method, which trains a SVM classifier with labeled data from both domains using the original bag-of-word features.

(2) *PLSA*: This is an unsupervised word clustering method, which first applies the probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) to obtain word clusterings with both labeled and unlabeled data from the two domains and then uses the soft word clusterings as augmenting features to train SVM classifiers.

(3) *FDLDA*: This is an alternative supervised word clustering method we built by training the Fast-Discriminative Latent Dirichlet Allocation model (Shan et al., 2009) with all labeled data from the two domains. After training the model, we used the learned topic distribution $p(z)$ and the conditional word distributions $p(w|z)$ to compute the conditional distribution over topics $p(z|w)$ for each word as the soft clustering of the word. We then used the soft word clusterings as augmenting features to train SVM classifiers.

(4) *SCL*: This is the structural correspondence learning based domain adaptation method (Blitzer et al., 2006). It first induces generalizable features with all data from both domains by modeling the correlations between pivot features and non-pivot features, and then uses the produced generalizable features as augmenting features to train SVM classifiers.

(5) *CPSP*: This is coupled subspace learning based domain adaptation method (Blitzer et al., 2011). It first learns two domain projectors using all data from the two domains by approximating multi-view dimensionality reduction, and then projects the labeled data to low dimensional latent feature space to train SVM Classifiers.

We used the LIBSVM package (Chang and Lin, 2011) with its default parameter setting to train linear SVM classifiers as the base classification model for all comparison methods.

Table 1: Average results (accuracy±standard deviation) for three cross-domain document categorization tasks on Reuters-21578 dataset.

Task	BOW	PLSA	FDLDA	SCL	CPSP	SWC
Orgs vs People	76.07±0.39	76.50±0.10	76.95±0.23	78.71±0.20	77.58±0.21	**81.27±0.23**
Orgs vs Places	73.88±0.58	74.68±0.20	74.87±0.29	76.71±0.23	75.76±0.28	**78.33±0.64**
People vs Places	61.80±0.44	63.36±0.40	63.46±0.40	64.65±0.40	62.73±0.53	**67.48±0.20**

4.2 Experiments on Reuters Data Set

We used the popularly studied Reuters-21578 dataset (Dai et al., 2007), which contains three cross-domain document categorization tasks, *Orgs vs People, Orgs vs Places, People vs Places*. The source and target domains of each task contain documents sampled from different non-overlapping subcategories. From example, the task of *Orgs vs People* assigns a document into one of the two top categories (*Orgs, People*), and the source domain documents and the target domain documents are sampled from different subcategories of *Orgs* and *People*. There are 1237 source documents and 1208 target documents for the task of *Orgs vs People*, 1016 source documents and 1043 target documents for the task of *Orgs vs Places*, and 1077 source documents and 1077 target documents for the task of*People vs Places*. For each task, we built a unigram vocabulary based on all the documents from the two domains and represented each document as a feature vector containing term frequency values.

4.2.1 Experimental Results for Cross-Domain Document Categorization

For each of the three cross-domain document categorization tasks on Reuters-21578 dataset, we used all the source documents as labeled training data while randomly selecting 100 target documents as labeled training data and setting the rest as unlabeled test data. For the BOW baseline method, we used the term-frequency features. The other five approaches are based on representation learning, and we selected the dimension size of the representation learning, i.e., the cluster number in our proposed approach, from $\{5, 10, 20, 50, 100\}$ according to the average classification results over 3 runs on the task of *Orgs vs People*. The dimension sizes of the induced representations for the five approaches, *PLSA, FDLDA, SCL, CPSP* and *SWC* are *20, 20, 100, 100* and *20* respectively.

We then repeated each experiment 10 times on each task with different random selections of the 100 labeled target documents to compare the six comparison approaches. The average classification results in terms of accuracy and standard deviations are reported in Table 1. We can see that by simply combining labeled documents from the two domains without adaptation, the *BOW* method performs poorly across the three tasks. The *PLSA* method outperforms the *BOW* method over all the three tasks with small improvements. The supervised word clustering method *FDLDA*, though performing slightly better than the unsupervised clustering method *PLSA*, produces poor performance comparing to the proposed *SWC* method. One possible reason is that the *FDLDA* model is not specialized for supervised word clustering, and it uses a logistic regression model to predict the labels from the word topics, while the final soft word clustering is computed from the learned distribution $p(z)$ and $p(w|z)$. That is, in the FDLDA model the labels only influence the word clusterings indirectly and hence its influence can be much smaller than the influence of labels as direct parent variables of the word cluster variables in the *SWC* model. The two domain adaptation approaches, *SCL* and *CPSP*, both produce significant improvements over *BOW*, *PLSA* and *FDLDA* on the two tasks of *Orgs vs People* and *Orgs vs Places*, while the *CPSP* method produces slightly inferior performance than *PLSA* and *FDLDA* on the task of *People vs Places*. The proposed method *SWC* on the other hand consistently and significantly outperforms all the other comparison methods across all the three tasks.

We also studied the sensitivity of the proposed approach with respect to the number of clusters,

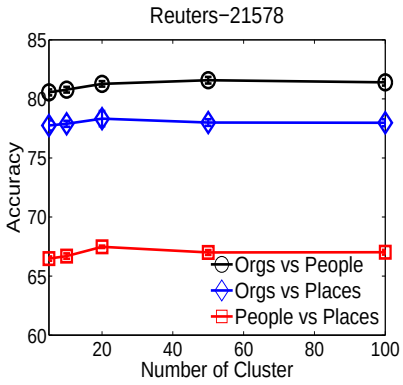

Figure 2: Sensitivity analysis of the proposed approach w.r.t. the number of clusters for the three cross-domain document categorization tasks on Reuters-21578 dataset.

i.e., the dimension size of the learned representation. We experimented with a set of different values $m \in \{5, 10, 20, 50, 100\}$ as the number of clusters. For each m value, we used the same experimental setting as above and repeated the experiments 10 times to obtain the average comparison results. The classification accuracy results on the three tasks are reported in Figure 2. We can see that the proposed method is not very sensitive to the number of clusters, across the set of increasing values we considered, and its performance becomes very stable after the cluster number reaches 20.

4.2.2 Document Categorization Accuracy vs Label Complexity in Target Domain

We next conducted experiments to compare the six approaches by varying the amount of the labeled data from the target domain. We tested a set of different values, $s \in \{100, 200, 300, 400, 500\}$, as the number of labeled documents from the target domain. For each different s value, we repeated the experiments 10 times by randomly selecting s labeled documents from the target domain using the same experimental setting as before. The comparison results across the set of s values are plotted in Figure 3. We can see that in general the performance of each method improves with the increase of the number of labeled documents from the target domain. The proposed method *SWC* and the domain adaptation method *SCL* clearly outperform the other four methods. Moreover, the proposed method *SWC* not only maintains consistent and significant advantages over all other methods across the range of different s values, its performance with 300 labeled target instances is even superior to the other methods with 500 labeled target instances. All these results suggest the proposed approach is very effective for adapting data across domains.

4.3 Experiments on Amazon Product Reviews

We conducted cross-domain sentiment classification on the widely used Amazon product reviews (Blitzer et al., 2007), which contains review documents distributed in four categories: *Books(B)*, *DVD(D)*, *Electronics(E)* and *Kitchen(K)*. Each category contains 1000 positive and 1000 negative reviews. We constructed 12 cross-domain sentiment classification tasks, one for each source-target domain pair, *B2D*, *B2E*, *B2K*, *D2B*, *D2E*, *D2K*, *E2B*, *E2D*, *E2K*, *K2B*, *K2D*, *K2E*. For example, the task *B2D* means that we use the *Books* reviews as the source domain and the *DVD* reviews as the target domain. For each pair of domains, we built a vocabulary with both unigram and bigram features extracted from all the documents of the two domains, and then represented each review document as a feature vector with term frequency values.

4.3.1 Experimental Results for Cross-Domain Sentiment Classification

For each of the twelve cross-domain sentiment classification tasks on Amazon product reviews, we used all the source reviews as labeled data and randomly selected 100 target reviews as labeled data while treating the rest as unlabeled test data. For the baseline method *BOW*, we used binary indicator values as features, which has been shown to work better than the term-frequency features for sentiment classification tasks (Pang et al., 2002; Na et al., 2004). For all the other representation learning based methods, we selected the dimension size of learned representation according to the average results over 3 runs on the *B2D* task. The dimension sizes selected for the methods *PLSA*, *FDLDA*, *SCL*, *CPSP*, and *SWC* are 10, 50, 50, 100 and 10, respectively.[1]

[1] 50 and 100 are also the suggested values for SCL (Blitzer et al., 2007) and CPSP (Blitzer et al., 2011) respectively on this cross-domain sentiment classification dataset.

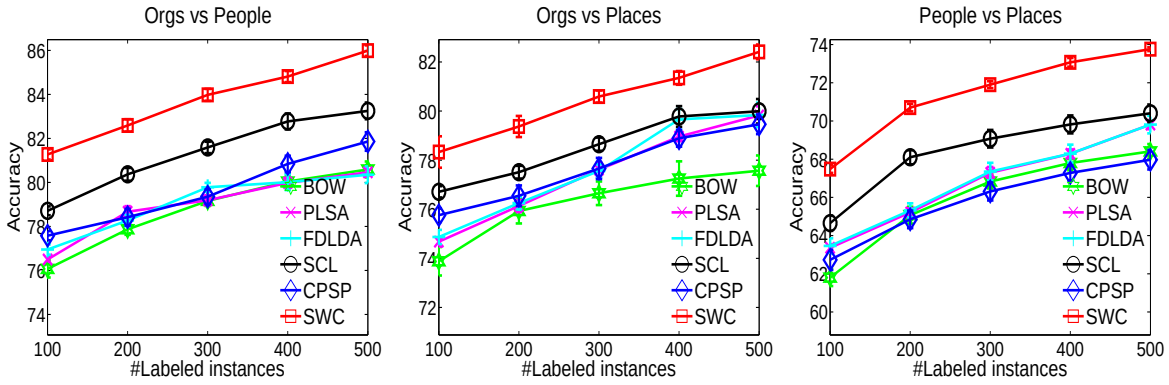

Figure 3: Average classification results for three cross-domain document categorization tasks on Reuters-21578 dataset by varying the amount of labeled training data from the target domain.

Table 2: Average results (accuracy±standard deviation) for twelve cross-domain sentiment classification tasks on Amazon product reviews.

Task	**BOW**	**PLSA**	**FDLDA**	**SCL**	**CPSP**	**SWC**
B2D	76.58±0.14	76.01±0.10	75.95±0.16	80.17±0.16	77.53±0.14	**81.66**±0.23
B2K	75.48±0.34	74.68±0.20	74.87±0.15	78.13±0.21	76.38±0.15	**82.26**±0.20
B2E	72.92±0.37	73.36±0.19	73.46±0.21	74.79±0.19	73.31±0.17	**77.04**±0.64
D2B	74.10±0.29	74.04±0.20	74.08±0.18	78.73±0.23	77.07±0.15	**79.95**±0.25
D2K	75.19±0.33	75.37±0.31	75.44±0.31	76.98±0.19	76.77±0.10	**82.13**±0.20
D2E	73.01±0.34	74.21±0.30	74.09±0.31	75.69±0.25	73.83±0.21	**76.98**±0.54
E2B	67.58±0.24	68.48±0.15	68.44±0.17	70.21±0.16	70.47±0.16	**72.11**±0.46
E2D	70.15±0.27	70.16±0.23	70.06±0.22	72.83±0.25	71.76±0.20	**73.81**±0.59
E2K	82.23±0.12	82.24±0.18	82.26±0.19	84.69±0.11	81.31±0.14	**85.33**±0.16
K2B	70.67±0.18	72.18±0.21	72.18±0.16	73.91±0.21	72.18±0.19	**75.78**±0.55
K2D	71.51±0.26	72.00±0.18	72.05±0.19	74.82±0.26	72.59±0.18	**76.88**±0.49
K2E	80.81±0.12	80.39±0.18	80.46±0.18	82.96±0.11	80.81±0.14	**84.78**±0.19

We then repeated each experiment 10 times based on different random selections of 100 labeled reviews from the target domain to compare the six methods on the twelve tasks. The average classification results are reported in Table 2. We can see that the *PLSA* and *FDLDA* methods do not show much advantage over the baseline method *BOW*. *CPSP* performs better than *PLSA* and *BOW* on many of the twelve tasks, but with small advantages, while *SCL* outperforms *CPSP* on most tasks. The proposed method *SWC* however demonstrates a clear advantage over all the other methods and produces the best results on all the twelve tasks.

We also conducted sensitivity analysis over the proposed approach regarding the number of clusters on the twelve cross-domain sentiment classification tasks, by testing a set of cluster number values $m = \{5, 10, 20, 50, 100\}$. The average results are plotted in Figure 5. Similar as before, we can see the proposed approach has stable performance across the set of different cluster numbers. Moreover, these results also clearly show that domain adaptation is not a symmetric process, as we can see it is easier to conduct domain adaptation from the source domain *Books* to the target domain *Kitchen* (with an accuracy around 82%), but it is more difficult to make domain adaptation from the source domain *Kitchen* to the target domain *Books* (with an ac-

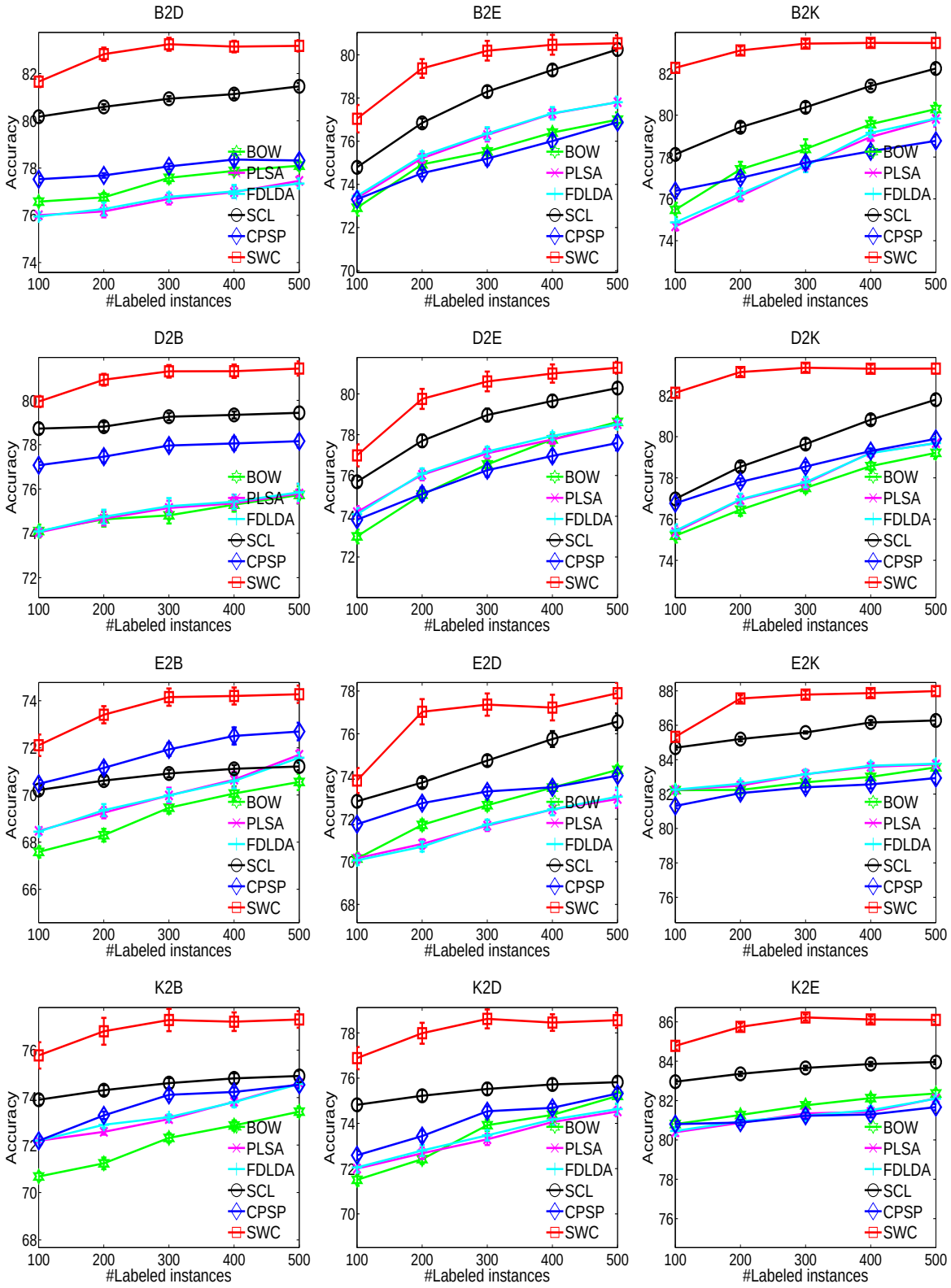

Figure 4: Average results (accuracy±standard deviation) for the 12 cross-domain sentiment classification tasks on Amazon product reviews with different numbers of labeled training data from the target domain.

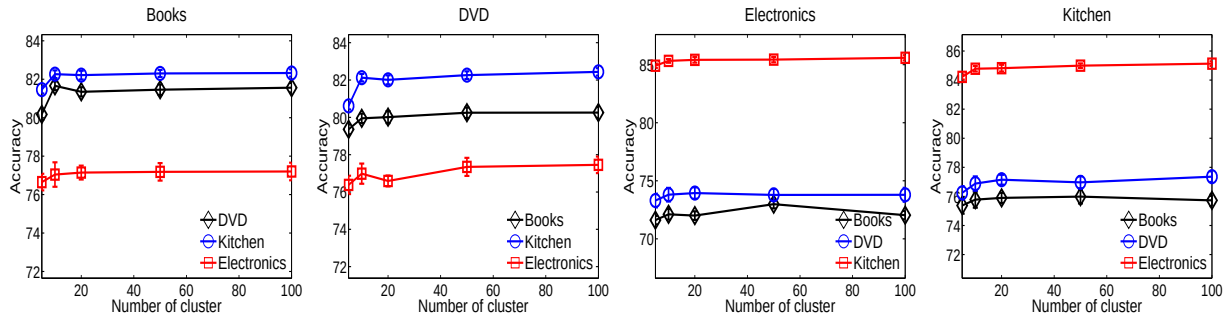

Figure 5: Sensitivity analysis of the proposed approach wrt the number of clusters for the twelve cross-domain sentiment classification tasks. Each figure shows experimental results for three tasks with the same source domain.

curacy around 75%). It also shows that the degree of relatedness of the two domains is an important factor for the effectiveness of knowledge adaptation. For example, one can see that it is much easier to conduct domain adaptation from *Kitchen* to *Electronics* (with an accuracy around 84%) than from *Kitchen* to *Books* (with an accuracy around 75%), as *Kitchen* is more closely related to *Electronics* than *Books*.

4.3.2 Sentiment Classification Accuracy vs Label Complexity in Target Domain

Similar as before, we tested the proposed approach using a set of different values $s \in \{100, 200, 300, 400, 500\}$ as the number of labeled reviews from the target domain. For each given s value, we conducted the comparison experiments using the same setting above. The average results are reported in Figure 4. We can see that the performance of each approach in general improves with the increase of the number of labeled reviews from the target domain. The proposed approach maintains a clear advantage over all the other methods on all the twelve tasks across different label complexities. All those empirical results demonstrate the effectiveness of the proposed approach for cross-domain sentiment classification.

4.3.3 Illustration of the Word Clusters

Finally, we would also like to demonstrate the *hard* word clusters produced by the proposed supervised word clustering method. We assign a word into the cluster it most likely belongs to according to its soft clustering representation, such as $c^* = \arg\max_c P(c|w, \boldsymbol{\theta}^*)$. Table 3 presents the top representative words (i.e., the most frequent words) of the 10 word clusters produced on the task of *B2K*. We can see that the first three clusters (C1, C2, and C3) contain words with *positive* sentiment polarity in different degrees. The two clusters (C4 and C5) contain words used to express the degree of opinions. The next four clusters (C6, C7, C8, and C9) contain content words related to *Books* or *Kitchen*. The last cluster (C10) contains words of *negative* sentiment polarity. These results demonstrate that the proposed supervised word clustering can produce task meaningful word clusters and hence label-informative latent features, which justifies its effectiveness.

5 Conclusion

In this paper, we proposed a novel supervised representation learning method to tackle domain adaptation by inducing predictive latent features based on supervised word clustering. With the soft word clustering produced, we can transform all documents from the two domains into a unified low-dimensional feature space for effective training of cross-domain NLP prediction system. We conducted extensive experiments on cross-domain document categorization tasks on Reuters-21578 dataset and cross-domain sentiment classification tasks on Amazon product reviews. Our empirical results demonstrated the efficacy of the proposed approach.

References

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adapta-

Table 3: Clustering illustration for the task of *B2K* on Amazon product reviews.

C1	recommend excellent wonderful beautiful love powerful happy satisfied outstanding
C2	enjoyed fantastic glad i_liked nicely was_great benefits pleasure amazingly
C3	good_and made_me most_people ordered_this standards accurately check_out
C4	was_a kind_of basically is_only half_of first_of as_if and_still anything_about have_some
C5	ever may still going maybe either at_least of_all totally sort_of are_very
C6	life work machine size design bottom business picture hand hook gas sink turner shelves
C7	way coffee pan keep cooking maker heat job working children handle meet core wine
C8	people us world come fact man place stars during example went short bathroom apple price
C9	pot friends daily light fire tells knew holds keep_the continued meal hooked silver wind
C10	disappointed waste unfortunately worse poorly sorry weak not_worth stupid fails awful useless

tion. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learng*, 79(1-2):151–175, 2010.

D. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

J. Blitzer, D. Foster, and S. Kakade. Domain adaptation with coupled subspaces. In *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

M. Chen, K. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Advances in Neural Inform. Process. Systems (NIPS)*, 2011.

M. Chen, Z. Xu, K. Weinberger, and F. Sha. Marginalized denoising autoencoders for domain adaptation. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2012.

W. Dai, Q. Yang, G. Xue, and Y. Yu. Boosting for transfer learning. In *Proc. of the International Conf. on Machine Learning (ICML)*, 2007.

H. Daumé III. Frustratingly easy domain adaptation. In *Proc. of the Annual Meeting of the Association for Comput. Linguistics (ACL)*, 2007.

H. Daumé III, A. Kumar, and A. Saha. Co-regularization based semi-supervised domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society*, 39 (1):1–38, 1977.

J. Finkel and C. Manning. Hierarchical bayesian domain adaptation. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2009.

T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999.

J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

Y. Mejova and P. Srinivasan. Crossing media streams with sentiment: Domain adaptation in

blogs, reviews and twitter. In *Proc. of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

J. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In *Proc. of the Conf. of the Inter. Society for Knowledge Organization*, 2004.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proc. of the Annual Meeting of the Association for Comput. Linguistics (ACL)*, 2010.

H. Shan, A. Banerjee, and N. Oza. Discriminative mixed-membership models. In *Proc. of the IEEE Inter. Conference on Data Mining (ICDM)*, 2009.

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

S. Tan. Improving scl model for sentiment-transfer learning. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2009.

C. Wan, R. Pan, and J. Li. Bi-weighting domain adaptation for cross-language text classification. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.