Transductive Representation Learning for Cross-Lingual Text Classification

Yuhong Guo and Min Xiao
Dept. of Computer and Information Sciences
Temple University
Philadelphia, PA, USA
{*yuhong, minxiao*}*@temple.edu*

***Abstract*—In cross-lingual text classification problems, it is costly and time-consuming to annotate documents for each individual language. To avoid the expensive re-labeling process, domain adaptation techniques can be applied to adapt a learning system trained in one language domain to another language domain. In this paper we develop a transductive subspace representation learning method to address domain adaptation for cross-lingual text classifications. The proposed approach is formulated as a nonnegative matrix factorization problem and solved using an iterative optimization procedure. Our empirical study on cross-lingual text classification tasks shows the proposed approach consistently outperforms a number of comparison methods.**

***Keywords*-domain adaptation; cross-lingual text classification; representation learning.**

I. Introduction

In cross-lingual text classification (CLTC) [1], it is costly and time-consuming to annotate documents for each individual language. To avoid the expensive re-labeling process, it is important to effectively adapt a learning system trained in one language (the source domain), which has plenty of labeled data, for another language (the target domain), where labeled data are scarce. Previous work on cross-lingual text classification tasks mainly relied on machine translations [2], [3], [4], since the two domains have different feature representation spaces. However, even with translated documents, domain divergence still exists in the form of feature distribution divergence due to the differences in culture, linguistic expression, and people's interest in different language regions. Domain adaptation techniques thus can be applied to address cross-language adaptations [3], [5], [6].

A major issue in domain adaptation is to minimize domain divergence to facilitate knowledge transfer from source domain to target domain. Many domain adaptation methods, including covariate shift methods [7], [8], [9] and feature representation learning methods, have been developed to bridge the domain divergence gap. It has been shown in a recent theoretical work that a good cross-domain feature representation is crucial for effective domain adaptation [10]. Previous work [11] demonstrates that a representation for successful domain adaptation must simultaneously maintain low values for both training error and domain divergence.

Recently, several approaches have been proposed to learn generalizable features from the source and target domains, including structural correspondence learning (SCL) [12], coupled subspace learning (CSL) [13], and feature augmentation methods [14], [15]. They all demonstrated good empirical performance for a variety of cross domain prediction tasks. However, the SCL and CSL methods reduce domain divergence and minimize training error sequentially instead of simultaneously. This inevitably leads to suboptimal representations for the given prediction task since the induced latent features do not convey task specific discriminative information. The EASYADAPT (EA) [14] and the co-regularization based semi-supervised extension of EA (EA++) [15], on the other hand, simultaneously minimize both domain divergence and training error. They however try to separate common features across domains from domain-specific features in the original feature space, instead of inducing common predictive feature representations. Moreover, all these approaches are unsuitable for domain adaptation tasks where the feature spaces of the two domains are different. For cross-lingual text classification tasks, although we can still employ these methods on the top of machine translation tools, they nevertheless suffer from the information loss and translation errors introduced in the machine translation process.

Multi-view learning methods in combination with machine translation have also been applied on cross-lingual text classification tasks recently, including co-training [16], multi-view majority voting [1], and multi-view co-regularization [17]. These method can exploit original documents in both the source and target domains, and thus can alleviate the problem of information loss and translation error. However, they mostly work on the original feature space of each language, without effectively addressing the feature distribution divergence problem between the original and translated documents in each language.

In this paper, we generalize one primary assumption of covariate shift methods for domain adaptation to tackle cross-lingual text classifications. We propose a transductive subspace representation learning method for domain adaptation with heterogeneous cross-domain feature spaces. We assume there is a shared latent subspace representation over the two domains, such that one common prediction function can be learned from the shared representation for both domains. Identifying such a common latent representation will automatically bridge the domain divergence gap. Our approach requires a translation tool to translate instances from one domain into the other domain. For cross-lingual text classification tasks, we use machine translation to translate texts across languages. Thus each text article is represented by two parallel instances, as two views of it. We then formulate the transductive subspace representation learning and the common prediction function training as one joint optimization problem over two-view data, which simultaneously minimizes both domain divergence and training error. The formulated problem is a special transductive nonnegative matrix factorization (NMF) problem, which not only can exploit information from two domains, but also can exploit information from both labeled and unlabeled instances. We develop an iterative optimization procedure to solve the problem for a local optimal solution. Our empirical results on cross-lingual text classification tasks suggest the proposed approach consistently outperforms several comparison methods.

II. Main Approach

In this section, we present a novel latent subspace representation learning method to conduct domain adaptation for domains with different feature spaces.

A. Motivation

A fundamental assumption behind many standard domain adaptation methods is that the two domains share the same conditional distribution, i.e., $P_S(Y|X=\mathbf{x}) = P_T(Y|X=\mathbf{x})$, but the marginal distributions vary across domains, $P_S(X=\mathbf{x}) \neq P_T(X=\mathbf{x})$. The problem of the marginal distribution varying has been widely studied as covariate shift [7], [8], [9]. However, the assumption of a common conditional distribution is obviously very constrictive and cannot hold when the two domains have different feature spaces. We propose to generalize this assumption to address heterogeneous feature spaces based on their common latent subspace representations. Specifically, we assume there is a shared latent subspace representation space $\mathcal{Z}$ instead of a shared conditional distribution over the two domains, such that one common prediction function, $f : \mathcal{Z} \to \mathcal{Y}$ (or a conditional distribution $P(Y|Z)$), can be learned for both domains. Note a latent subspace representation vector $\mathbf{z}$ can correspond to different feature representation vectors in different domains. Thus the generalized latent representation based assumption holds for domain adaptations with either homogeneous or heterogeneous feature spaces. It is obvious that identifying such common latent representations will automatically bridge the domain convergence gap.

Like many domain adaptation methods developed for cross-lingual text classifications where two domains have different feature spaces, we require the assistance of machine translation tools. We first use machine translation tools to translate each instance in one domain into a parallel instance in the other domain. Thus each text article is represented by two instances as two views. Different from many multi-view learning methods exploited in cross-lingual text classifications, we try to capture the deep underlying predictive data structure shared by the two views. We formulate the common latent subspace representation learning and the common prediction function training as one joint optimization problem in form of nonnegative matrix factorization, which simultaneously minimizes both domain divergence and training error. Such an optimization objective is supported by previous theoretical work in domain adaptation [11].

B. Notation and Setting

Let $X_s = [X_s^\ell; X_s^u] \in \mathbb{R}_+^{n_s \times d_s}$ denote the nonnegative data matrix from the source domain, where X_s^ℓ is the labeled submatrix and X_s^u is the unlabeled submatrix. Let $\mathbf{y}_s^\ell \in \{+1, -1\}^{l_s}$ denote the corresponding label vector. Similarly, let $X_t = [X_t^\ell; X_t^u] \in \mathbb{R}_+^{n_t \times d_t}$ denote the nonnegative data from the target domain, where X_t^ℓ is the labeled submatrix and X_t^u is the unlabeled submatrix, and $\mathbf{y}_t^\ell \in \{+1, -1\}^{l_t}$ denote the corresponding label vector. Typically we assume there are much more labeled instances in the source domain than in the target domain, i.e., $l_s > l_t$.

For the source data matrix X_s, we can construct its parallel view $\widehat{X}_s$ in the target domain using a transformation tool. Similarly, for X_t we can construct its parallel view $\widehat{X}_t$ in the source domain. Collecting everything together, we have parallel data in two views: $X_1 = [X_s^\ell; \widehat{X}_t^\ell; X_s^u; \widehat{X}_t^u] \in \mathbb{R}_+^{n \times d_s}$, and $X_2 = [\widehat{X}_s^\ell; X_t^\ell; \widehat{X}_s^u; X_t^u] \in \mathbb{R}_+^{n \times d_t}$, with labels $\mathbf{y}^\ell = [\mathbf{y}_s^\ell; \mathbf{y}_t^\ell] \in \{+1, -1\}^l$.

C. Predictive Latent Subspace Representation Learning with Non-Negative Matrix Factorization

To address the cross-lingual text classification problem, where the data in two languages are both represented using nonnegative features, but in different feature spaces, we extend the nonnegative matrix factorization [18] in two ways: discovering common low-dimensional representations from parallel data expressed in different feature spaces, and incorporating discriminative label information.

Given parallel data matrices X_1 and X_2, we aim to find a common latent subspace representation matrix Z for them such that the original data information can be maximally preserved by enforcing $X_1 \approx Z\Theta_1$ and $X_2 \approx Z\Theta_2$ for nonnegative subspace matrices Θ_1 and Θ_2, while enabling a good prediction model to be learned from Z with minimal training error.

Let Z^ℓ denotes the subset of Z corresponding to the labeled part of X_1 and X_2, i.e., the first l rows of X_1 and X_2. For a latent vector $\mathbf{z}$, we assume a linear prediction function

$$f(\mathbf{z}) = \mathbf{z}^\top \mathbf{w} + b \tag{1}$$

where $\mathbf{w}$ is a $m \times 1$ parameter vector, and b is a bias parameter. Using a least squared loss function we formulate a joint learning problem as below

$$\begin{aligned}\min_{\mathbf{w},b,Z,\Theta_1,\Theta_2} \quad & \|\mathbf{y}^\ell - Z^\ell \mathbf{w} - b\mathbf{1}\|^2 + \beta\|\mathbf{w}\|^2 \qquad (2)\\ & +\alpha_1\|X_1 - Z\Theta_1\|_F^2 + \alpha_2\|X_2 - Z\Theta_2\|_F^2\\ \text{subject to} \quad & \Theta_1 \geq 0,\ \Theta_2 \geq 0,\ Z \geq 0\end{aligned}$$

where Z is a $n \times m$ nonnegative latent matrix, Θ_1 is a $m \times d_s$ matrix, and Θ_2 is a $m \times d_t$ matrix; $\mathbf{1}$ denotes a $l \times 1$ vector with all 1 entries; $\|\cdot\|_F$ denotes a Frobenius matrix norm. This objective function contains three components: a L2-norm regularized least squared training loss, and two matrix reconstruction losses. We expect this minimization problem will capture the intrinsic structure of the two view data and provide task-specific discriminative representation for learning the target prediction model.

First, we solve the minimization problem with respect to $\mathbf{w}$ and b given other variables fixed. By setting the derivatives of the objective function with respect to $\mathbf{w}$ and b to zeros, we obtain the following closed-form solutions

$$\begin{aligned} b &= \frac{1}{l}\mathbf{1}^\top(\mathbf{y}^\ell - Z^\ell\mathbf{w}) \qquad (3)\\ \mathbf{w} &= (Z^{l\top}HZ^\ell + \beta I)^{-1}Z^{l\top}H\mathbf{y}^\ell \qquad (4)\end{aligned}$$

where $H = I - \frac{1}{l}\mathbf{1}\mathbf{1}^\top$ is a $l \times l$ matrix, and I is a $l \times l$ identity matrix. Plugging (3) and (4) back into the objective function, we obtain the following optimization problem

$$\begin{aligned}\min_{Z,\Theta_1,\Theta_2} \quad & \alpha_1\|X_1 - Z\Theta_1\|_F^2 + \alpha_2\|X_2 - Z\Theta_2\|_F^2 \quad (5)\\ & -\mathbf{y}^{l\top}HZ^\ell(Z^{l\top}HZ^\ell + \beta I)^{-1}Z^{l\top}H\mathbf{y}^\ell\\ \text{subject to} \quad & \Theta_1 \geq 0,\ \Theta_2 \geq 0,\ Z \geq 0.\end{aligned}$$

Let $B = [I, O_{l,u}]$ where $O_{l,u}$ is a $l \times u$ matrix with all zero values, such that $Z^\ell = BZ$. Let $A = HB$. Then the optimization problem (5) can be rewritten as

$$\begin{aligned}\min_{Z,\Theta_1,\Theta_2} \quad & L = \alpha_1\|X_1 - Z\Theta_1\|_F^2 + \alpha_2\|X_2 - Z\Theta_2\|_F^2\\ & -\mathbf{y}^{l\top}AZ(Z^\top A^\top AZ + \beta I)^{-1}Z^\top A^\top\mathbf{y}^\ell \quad (6)\\ \text{subject to} \quad & \Theta_1 \geq 0,\ \Theta_2 \geq 0,\ Z \geq 0.\end{aligned}$$

The objective function of the semi-supervised optimization problem above has two matrix factorization terms, one for each view. The third term takes the discriminative label information from both domains into account, which only involves the small set of labeled data and has a matrix inversion term over a $m \times m$ matrix. Later we will show in our experiments, m can be a small number, e.g., $m = 20$. Thus the computational complexity of the optimization problem above is not much higher than the standard NMFs.

III. Optimization Algorithm

The optimization problem formulated above can be viewed as a transductive nonnegative matrix factorization (NMF) problem. However, the simple multiplicative updates used in standard nonnegative matrix factorizations cannot be directly applied here due to the fact that our objective function is much more general with a matrix inverse term. We develop an *iterative projected gradient descent* optimization algorithm to solve it for a local optimal solution.

Let $L(Z, \Theta_1, \Theta_2)$ denote the objective function in (6). Starting from some initial feasible values $\{Z, \Theta_1, \Theta_2\}$, in each iteration we sequentially update each parameter matrix using a projected gradient descent method to minimize the objective function. To facilitate the algorithm presentation below, we define a general projection function

$$Proj(X; \mathcal{C}) = \arg\min_{M: M\in\mathcal{C}} \|M - X\|_F^2 \qquad (7)$$

which projects a given matrix X into the constraint set $\mathcal{C}$.

A. Update Z

In each iteration, given the current values of Z, Θ_1 and Θ_2, we update Z in the following way to improve the objective function L. First we compute the derivative of L with respect to Z,

$$\begin{aligned}\frac{\partial L}{\partial Z} = \; & 2\alpha_1 Z\Theta_1\Theta_1^\top - 2\alpha_1 X_1\Theta_1^\top + 2\alpha_2 Z\Theta_2\Theta_2^\top \quad (8)\\ & -2\alpha_2 X_2\Theta_2^\top - 2A^\top\mathbf{y}^\ell\mathbf{y}^{l\top}AZQ\\ & +2A^\top AZQZ^\top A^\top\mathbf{y}^\ell\mathbf{y}^{l\top}AZQ\end{aligned}$$

where $Q = (Z^\top A^\top AZ + \beta I)^{-1}$. Let $\mathcal{C}_Z = \{M : M \in \mathbb{R}^{n\times m}, M \geq 0\}$. We then update Z along its gradient direction and project the updated matrix into the feasible set $\mathcal{C}_Z$; i.e.,

$$Z = Proj\left(Z - \tau^*\frac{\partial L}{\partial Z}; \mathcal{C}_Z\right) \qquad (9)$$

where τ^* is an optimal stepsize parameter. The projection (9) has a simple closed-form solution $Z = \max(Z - \tau^*\frac{\partial L}{\partial Z}, 0)$. Moreover, following the principle of standard NMF [18], we ensure $\mathbf{1}^\top Z\mathbf{1} = 1$ to remove the scale factor between Z and Θ_1, Θ_2. This can be simply achieved by taking a default step $Z = Z/(\mathbf{1}^\top Z\mathbf{1})$ following each projection operation. The optimal stepsize τ^* can be determined using a backtracking line search procedure by minimizing the objective function; that is,

$$\tau^* = \arg\min_{0<\tau\leq 1} L\left(Proj\left(Z - \tau\frac{\partial L}{\partial Z}; \mathcal{C}_Z\right), \Theta_1, \Theta_2\right) \qquad (10)$$

Algorithm 1 A Projected Gradient Descent Algorithm

```
Input: parallel data matrices $X_1$ and $X_2$; labels $\mathbf{y}^\ell$;
       initial $Z$, $\Theta_1$ and $\Theta_2$;
       control parameters $\alpha_1, \alpha_2, \beta, \epsilon > 0$.
Procedure:
  for iter = 1 to maxiters do
    • Set $\widehat{Z} = Z, \widehat{\Theta}_1 = \Theta_1, \widehat{\Theta}_2 = \Theta_2$
    • Update $Z$ according to Eq. (9)
    • Set $Z = Z/(\mathbf{1}^\top Z\mathbf{1})$
    • Update $\Theta_1, \Theta_2$ according to Eq. (12)
    • Set $\sigma_z = \|\widehat{Z} - Z\|_F$, $\sigma_{j\in\{1,2\}} = \|\widehat{\Theta}_j - \Theta_j\|_F$,
        $\sigma_L = |L(\widehat{Z}, \widehat{\Theta}_1, \widehat{\Theta}_2) - L(Z, \Theta_1, \Theta_2)|$
    • if $(\sigma_z < \epsilon) \cdot (\sigma_1 < \epsilon) \cdot (\sigma_2 < \epsilon) \cdot (\sigma_L < \epsilon)$
      then break end if
  end for
```

B. *Update* $\{\Theta_1, \Theta_2\}$

Next, we update each Θ_i for $i \in \{1, 2\}$ using a projected gradient descent procedure as well. The derivative of L regarding Θ_i can be computed as

$$\frac{\partial L}{\partial \Theta_i} \;=\; 2\alpha_i Z^\top Z\Theta_i - 2\alpha_i Z^\top X_i \tag{11}$$

Let $\mathcal{C}_{\Theta_i} = \{M : M \in \mathbb{R}^{m\times d_i}, M \geq 0\}$, where $d_1 = d_s, d_2 = d_t$. Then Θ_i can be updated by

$$\Theta_i = Proj\left(\Theta_i - \tau^* \frac{\partial L}{\partial \Theta_i}; \mathcal{C}_{\Theta_i}\right) \tag{12}$$

where τ^* is an optimal stepsize parameter. Similar as above, a simple closed-form solution exists for (12), and the stepsize parameter τ^* can be found by using a backtracking line search to minimize the objective function.

We iteratively update Z, Θ_1, and Θ_2 as described above until local convergence is reached. The overall projected gradient descent algorithm is presented in Algorithm 1.

After solving for a local optimal solution $\{\Theta_1, \Theta_2, Z\}$, the prediction model parameters $\mathbf{w}$ and b can be recovered using equations (4) and (3) respectively. An unlabeled instance $\mathbf{z}^u$ can then be classified using $f(\mathbf{z}^u) = \mathbf{z}^{u\top}\mathbf{w} + b$.

IV. Experiments

In this section, we report our experimental results over cross lingual text classification tasks.

A. *Experimental Setting*

Our experiments are conducted on the cross-lingual text classification (CLTC) dataset used in [1]. This dataset is a comparable corpus constructed with data sampled from Reuters RCV1 and RCV2, which contain newswire articles written in 5 languages, namely, English(E), French(F), German(G), Italian(I) and Spanish(S), distributed over 6 classes. In this multilingual corpus, each original document was translated into the other 4 languages using a statistical machine translation system. The following preprocessing steps are conducted, including lowercasing all tokens, removing non-alphanumeric tokens, mapping digits to a single *digit* token, and filtering out the stopping words and tokens occurring in less than 5 documents. All documents are represented using TF-IDF features.

We aim to study cross-lingual text classification between multiple languages, especially between English the other 4 languages. Towards this goal, we constructed a set of binary cross-lingual classification tasks. We selected two classes, CCAT and ECAT, to form a binary classification problem. From all instances in these two classes, we randomly selected 2000 instances for each class in each language to use as our experimental data. From the 5 languages, we constructed 20 cross lingual binary classification tasks as shown in Figure I, one for each source-target pair of languages; for example, the task *E2F* means that the source domain is *English* and the target domain is *French*.

Approaches. We compared the following 7 cross domain classification approaches in our experiments. All of them used least squared predictors as base classifiers and were tested on the unlabeled instances in the target domain.

- **TVT**: a baseline approach that trains a classifier only on the labeled target instances.
- **TVST**: a baseline approach that trains a classifier on both the labeled target instances and the translated labeled instances from the source domain.
- **EA**: the EASYADAPT approach developed in [14]. It uses a synthetic source domain formed by translating the labeled data in the original source domain into the target language.
- **EA++**: the co-regularization based semi-supervised domain adaptation approach developed in [15]. It uses a synthetic source domain formed by translating all instances in the original source domain into the target language.
- **CSL**: the coupled subspace learning approach developed in [13]. It uses the same synthetic source domain as the EA++ above.
- **MVCC**: the semi-supervised version of the multi-view co-classification method [17], which penalizes the disagreement of the two view predictions on unlabeled data. It uses all parallel data in both source and target domains constructed by translating instances in each domain into the other domain.
- **NMF**: Our proposed domain adaptation approach.

B. *Classification Results*

For each of the 20 CLTC tasks constructed with different pairs of languages, we randomly chose 900 labeled and 100 unlabeled instances from the source domain, and chose 100 labeled and 900 unlabeled instances from the target domain to conduct each experiment. The translated parallel view of each instance in the alternative domain was produced using

Table I

AVERAGE ACCURACY±STANDARD DEVIATION RESULTS OVER 10 RUNS FOR 20 CLTC TASKS.

Tasks	**TVT**	**TVST**	**EA**	**EA++**	**CSL**	**MVCC**	**NMF**
E2F	79.00 ± 0.66	79.66 ± 0.63	78.33 ± 1.00	79.39 ± 0.66	73.93 ± 2.83	81.61 ± 0.56	$\mathbf{82.77} \pm 0.71$
E2G	76.49 ± 0.42	74.82 ± 0.63	75.18 ± 0.68	74.64 ± 0.66	75.76 ± 3.28	81.52 ± 0.59	$\mathbf{82.77} \pm 0.71$
E2I	80.37 ± 0.60	77.39 ± 0.72	79.97 ± 0.87	77.17 ± 0.79	76.14 ± 1.43	81.72 ± 0.77	$\mathbf{84.10} \pm 0.79$
E2S	85.63 ± 1.45	85.50 ± 0.90	84.17 ± 1.38	84.66 ± 0.99	78.52 ± 2.11	89.46 ± 0.46	$\mathbf{91.09} \pm 0.82$
F2E	77.14 ± 1.10	80.43 ± 0.53	77.84 ± 0.97	79.84 ± 0.53	73.49 ± 2.43	80.31 ± 0.64	$\mathbf{84.28} \pm 0.47$
F2G	77.56 ± 0.60	75.44 ± 0.85	77.78 ± 0.63	75.34 ± 0.77	81.64 ± 0.94	$\mathbf{82.60} \pm 0.69$	82.34 ± 0.51
F2I	80.37 ± 0.60	77.39 ± 0.72	76.48 ± 0.89	76.70 ± 0.79	76.14 ± 1.43	80.70 ± 0.48	$\mathbf{84.10} \pm 0.79$
F2S	85.44 ± 1.23	85.93 ± 0.84	83.21 ± 1.20	85.60 ± 0.54	76.39 ± 3.44	89.73 ± 0.50	$\mathbf{91.08} \pm 0.40$
G2E	76.64 ± 0.98	78.22 ± 0.78	76.14 ± 2.35	77.91 ± 0.74	77.96 ± 1.16	78.76 ± 0.67	$\mathbf{79.87} \pm 0.57$
G2F	76.06 ± 0.53	77.09 ± 0.56	76.93 ± 0.43	76.99 ± 0.55	77.89 ± 0.87	78.96 ± 0.51	$\mathbf{80.97} \pm 0.67$
G2I	79.94 ± 0.73	78.96 ± 0.89	79.40 ± 1.12	79.18 ± 0.76	75.48 ± 3.13	80.94 ± 0.75	$\mathbf{81.19} \pm 0.65$
G2S	86.59 ± 1.03	86.04 ± 0.45	85.59 ± 0.63	85.49 ± 0.46	62.26 ± 2.83	86.76 ± 0.44	$\mathbf{89.23} \pm 1.87$
I2E	77.11 ± 0.93	76.77 ± 0.60	$\mathbf{78.93} \pm 0.67$	76.48 ± 0.58	70.41 ± 2.61	77.42 ± 0.86	78.58 ± 0.70
I2F	77.07 ± 0.48	78.40 ± 0.65	77.84 ± 0.69	78.19 ± 0.65	74.01 ± 2.57	78.63 ± 0.53	$\mathbf{79.81} \pm 0.53$
I2G	77.58 ± 0.51	77.37 ± 0.41	78.69 ± 0.40	77.38 ± 0.27	76.72 ± 2.34	80.64 ± 0.36	$\mathbf{82.36} \pm 0.57$
I2S	86.87 ± 1.13	87.74 ± 0.37	85.26 ± 1.10	87.09 ± 0.41	75.98 ± 2.48	89.86 ± 0.46	$\mathbf{90.11} \pm 0.51$
S2E	74.29 ± 1.05	74.84 ± 0.51	69.44 ± 2.89	75.29 ± 0.48	74.17 ± 1.18	$\mathbf{78.98} \pm 0.45$	78.90 ± 0.81
S2F	76.22 ± 0.56	77.21 ± 0.59	76.39 ± 0.57	77.39 ± 0.59	75.59 ± 2.55	79.74 ± 0.41	$\mathbf{81.42} \pm 0.46$
S2G	76.86 ± 0.44	74.10 ± 0.80	73.58 ± 0.75	74.24 ± 0.64	77.83 ± 1.02	80.31 ± 0.53	$\mathbf{80.49} \pm 0.95$
S2I	79.22 ± 1.02	78.32 ± 0.78	80.04 ± 0.65	78.54 ± 0.90	73.41 ± 2.84	83.87 ± 0.47	$\mathbf{85.56} \pm 0.69$

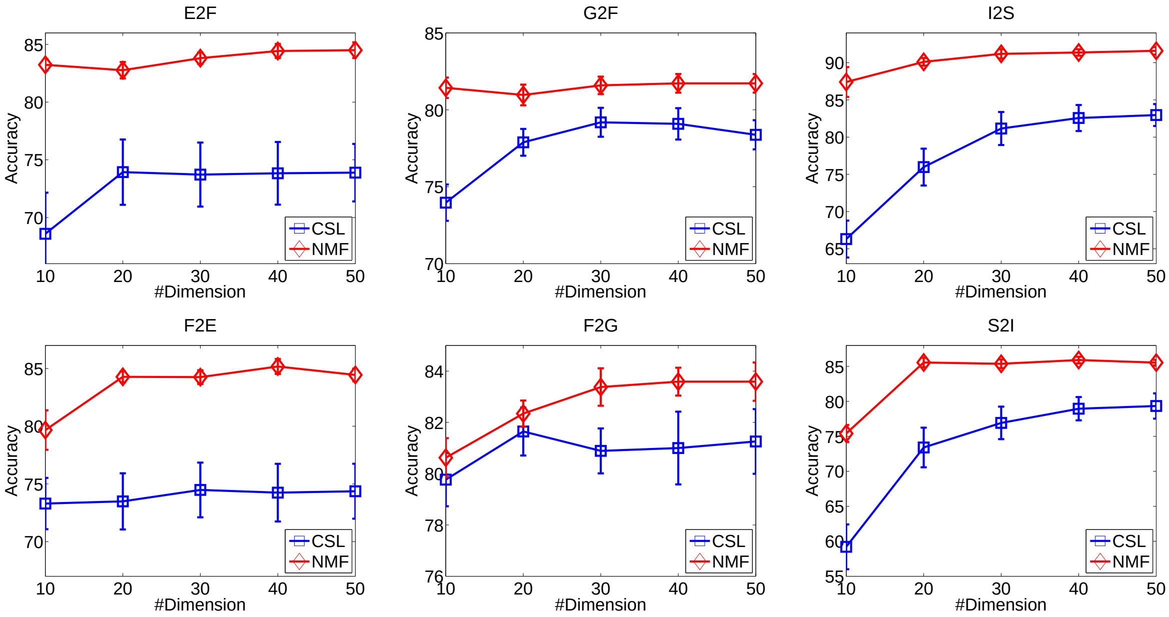

Figure 1. Classification accuracy results with different latent dimensions for 6 CLTC tasks.

machine translation. Thus finally each of the two view data matrices contains 2000 documents. For the two approaches, CSL and NMF, that learn latent feature representations, we set the dimension of the latent features as 20. The proposed NMF approach is not sensitive to the L_2 regularization parameter β, and we treated the two parallel views in a similar way. Thus we used $\alpha_1 = 0.5, \alpha_2 = 0.5, \beta = 0.0001$ for NMF. We repeated each experiment 10 times based on different random selections of source and target instances. The average test results on unlabeled (during training) target data, measured in term of accuracy, are reported in Table I.

From Table I, we can see that by simply exploiting translated labeled source data, TVST does not consistently outperform TVT. Comparing to the simple monolingual methods, TVT and TVST, the domain adaptation methods, EA, EA++ and CSL, lead to improved performances on some of the 20 tasks, but there are no consistent advantages. The multi-view method (MVCC), which exploits original data and translated data in both domains, demonstrates superior performance on most tasks, comparing to the baseline and domain adaptation methods. The proposed NMF, which combines multi-view learning and domain adaptation in one model, makes further advances. It outperforms the domain adaptation methods on 19 out of the 20 tasks, and outperforms the multi-view method on 18 out of the 20 tasks. These results suggest that domain divergence exists between the translated data and the original data in the same feature space, and machine translation is far from ideal in CLTC tasks. Thus the proposed technique that exploits the original data in both domains to alleviate the information loss of machine translation and learns a common subspace representation from two domains has advantages over methods that does not properly consider either original data or domain adaptation.

C. *Robustness to Latent Dimensionality*

Next we empirically studied the influence of latent dimension size over the two representation learning based methods, CSL and NMF, on a few tasks. We used the same data and experimental setting stated above for a set of different latent dimension sizes, $m = \{10, 20, 30, 40, 50\}$. The average classification accuracy results are presented in Figure 1. We can see the proposed approach NMF is less sensitive to latent dimension size and it is more effective than CSL across the range of latent dimension sizes.

V. Conclusion

In this paper we generalized the covariate shift assumption into heterogeneous cross-domain feature spaces by assuming a latent low dimensional representation is shared by both domains. We developed a novel predictive latent subspace representation learning approach to address domain adaptation problems with different feature spaces, in particular cross-lingual text classifications. The proposed approach conducts latent subspace representation learning and common prediction function training jointly in one nonnegative matrix factorization (NMF) problem, which simultaneously minimizes both domain divergence and training error. We developed an iterative optimization procedure to solve the problem for a local optimal solution. Our empirical results on cross-lingual text classification tasks suggest the proposed approach consistently outperforms a few baseline methods and some existing multi-view learning and domain adaptation methods.

References

[1] M. Amini, N. Usunier, and C. Goutte, “Learning from multiple partially observed views - an application to multilingual text categorization,” in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[2] J. G. Shanahan, G. Grefenstette, Y. Qu, and D. A. Evans, “Mining multilingual opinions through classification and translation,” in *Proc. of AAAI’04 Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

[3] L. Shi, R. Mihalcea, and M. Tian, “Cross language text classification by model translation and semi-supervised learning,” in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

[4] L. Rigutini and M. Maggini, “An EM based training algorithm for cross-language text categorization,” in *Proc. of the Web Intelligence Conference*, 2005.

[5] B. Wei and C. Pal, “Cross lingual adaptation: an experiment on sentiment classifications,” in *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

[6] P. Prettenhofer and B. Stein, “Cross-language text classification using structural correspondence learning,” in *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

[7] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[8] S. Bickel and T. Scheffer, “Dirichlet-enhanced spam filtering based on biased samples,” in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[9] 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)*, 2008.

[10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan, “A theory of learning from different domains,” *Machine Learning*, vol. 79, pp. 151–175, 2010.

[11] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[12] 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.

[13] 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.

[14] H. Daumé III, “Frustratingly easy domain adaptation,” in *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

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

[16] X. Wan, “Co-training for cross-lingual sentiment classification,” in *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2009.

[17] M. Amini and C. Goutte, “A co-classification approach to learning from multilingual corpora,” *Machine Learning*, vol. 79, pp. 105–121, 2010.

[18] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems (NIPS)*, 2000.