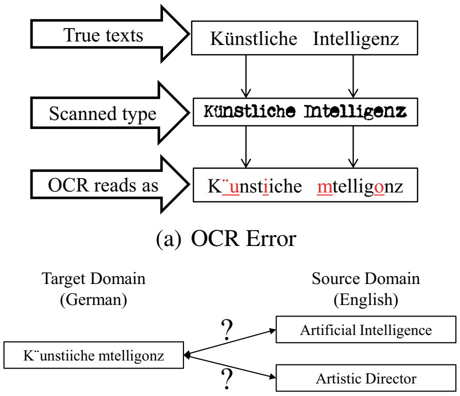
Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

# Transfer Learning for Cross-Language Text Categorization through Active Correspondences Construction*

**Joey Tianyi Zhou**[†], **Sinno Jialin Pan**[‡], **Ivor W. Tsang**[§], and **Shen-Shyang Ho**[‡]
[†]Institute of High Performance Computing, Singapore
[‡]Nanyang Technological University, Singapore
[§]QCIS, University of Technology Sydney, Australia
tzhou1@e.ntu.edu.sg, sinnopan@ntu.edu.sg, ivor.tsang@uts.edu.au, ssho@ntu.edu.sg

## Abstract

Most existing heterogeneous transfer learning (HTL) methods for cross-language text classification rely on sufficient cross-domain instance correspondences to learn a mapping across heterogeneous feature spaces, and assume that such correspondences are given in advance. However, in practice, correspondences between domains are usually unknown. In this case, extensively manual efforts are required to establish accurate correspondences across multilingual documents based on their content and meta-information. In this paper, we present a general framework to integrate active learning to construct correspondences between heterogeneous domains for HTL, namely HTL through active correspondences construction (HTLA). Based on this framework, we develop a new HTL method. On top of the new HTL method, we further propose a strategy to actively construct correspondences between domains. Extensive experiments are conducted on various multilingual text classification tasks to verify the effectiveness of HTLA.



(a) OCR Error

(b) Uncertainty on Finding Correspondences

Figure 1: Illustration of Motivation

## Introduction

In natural language processing, how to exploit knowledge in a source-language domain with sufficient labeled documents to assist text mining in a target-language domain with limited annotated documents is crucial. Recently, heterogeneous transfer learning (HTL) has been proposed to solve this problem (Duan, Xu, and Tsang 2012; Xiao and Guo 2013; Zhou et al. 2014a). One of the critical issues in HTL is how to learn a mapping between heterogeneous feature spaces of the source and target domains. Most existing HTL approaches assume that sufficient instance-correspondences are given in advance to learn such a feature mapping (Xiao and Guo 2013; Zhou et al. 2014a). However, in practice, instance-correspondences between domains are usually unknown. Therefore, extensively manual efforts are required to establish accurate correspondences across multilingual documents based on their content and meta-information.

For instance, given a task on topic categorization for scanned books in a minor language such as German or Vietnamese, there may be few annotated books available because manually constructing a precise hierarchy of categories is expensive, especially for a minor language. In contrast, there are plenty of annotated books in a major language, such as English. Intuitively, one can apply HTL techniques to transfer annotation information from a major language to a minor language for book categorization. This usually requires a sufficient set of correspondences between languages as inputs. However, due to the optical character recognition (OCR) error caused by converting the scanned books into editable data, simply machine/human translation on book titles fails to construct precise book correspondences between languages. For example, as shown in Figure 1(a), the book in German entitled “Künstliche Intelligenz” turns into unrecognized text “K¨unstiiche mtelligonz” by an OCR reader. Such error may induce ambiguity when constructing cross-language correspondences shown in Figure 1(b). The German book named “Künstliche Intelligenz” after OCR scanning and conversing can correspond to either the book “Artificial Intelligence” (“Künstliche Intelligenz” in German) or the book “Artistic Director” (“Künstlerisch Intelligenzia” in German) in English. In this case, extensive human efforts are required to construct the precise correspondences based on the content and meta information of the books.

Inspired by the idea of active learning (Settles 2010), to reduce the cost of constructing correspondences between

*This work was done when the first author was with Nanyang Technological University, Singapore.

domains for HTL, one can actively construct most "informative" and "high-quality" correspondences with limited budget for effective knowledge transfer across domains. In this paper, we propose a novel and general HTL framework named Heterogeneous Transfer Learning through Active correspondence construction (HTLA) to address the problem. In summary, our contributions are two folds:

- We present a novel framework to actively construct correspondences between heterogeneous domains for effective knowledge transfer.
- We present a new correspondence-based method for HTL through matrix completion with a regularization term on distribution matching. The encoding of the regularization into matrix completion significantly improves the prediction performance of HTL, but introduces additional challenges in optimization. We propose an effective approach to solve the optimization problem.

## Related Work

Different from homogeneous transfer learning, which aims to transfer knowledge across domains that are of homogeneous features (Pan and Yang 2010; Daumé III 2007), heterogeneous transfer learning (HTL) is proposed for heterogeneous domains with non-overlapping features. Most existing HTL methods focus on two settings. One assumes a few target-domain labeled data be available in training (Kulis, Saenko, and Darrell 2011; Duan, Xu, and Tsang 2012; Wang and Mahadevan 2011; Zhou et al. 2014b), while the other assumes sufficient instance-correspondences between heterogeneous domains be available. In the latter setting, which is our focus, Dai et al. (2008) proposed a probabilistic model to construct a "translator" to build connections between instances from different domains. Recently, Xiao and Guo (2013) applied an existing matrix completion technique to HTL. Specifically, in their proposed method, some instance-correspondences between domains are assumed to be given in advance, and the goal is to reconstruct "missing correspondences" for all the instances observed in either the source or target domain. However, their proposed method does not consider distribution distance between domains when reconstructing the corresponding instances. In this paper, we propose a more general matrix completion method for HTL by taking distribution distance minimization into consideration. Moreover, we propose a framework to actively construct instance-correspondences for HTL.

There has been research work on combining active learning and transfer learning to actively query instances for labels in the target domain by leveraging knowledge from the source domain (Shi, Fan, and Ren 2008; Rai et al. 2010; Chattopadhyay et al. 2013; Wang, Huang, and Schneider 2014). However, most of them are focused on homogeneous transfer learning. The most related work to ours on actively constructing correspondences between heterogeneous domains for knowledge transfer is (Zhao et al. 2013), where an active learning strategy was proposed to selectively identify user/item correspondences between different recommender systems. However, their proposed method is specific for collaborative filtering in recommender systems, which is not easy to generalize to other HTL problems. Our proposed algorithm is general for cross-language text classification.

## Problem Statement and Formulation

Assume that we are given a set of source-domain labeled instances, $\{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{l}$, where $\mathbf{x}_{S_i} \in R^{d_S \times 1}$ denotes an instance and $y_{S_i} \in \{1, -1\}$ denotes the corresponding label, a set of source-domain unlabeled instances $\{\mathbf{x}_{S_i}\}_{i=l+1}^{n_S}$, and a set of corresponding pairs between the source and target domains, $\{(\mathbf{x}_{S_i}^C, \mathbf{x}_{T_i}^C)\}_{i=1}^{n_C}$, where $\mathbf{x}_{S_i}^C \in R^{d_S \times 1}$ and $\mathbf{x}_{T_i}^C \in R^{d_T \times 1}$ are source- and target- domain instances respectively. Note that $n_C$ can be 0 initially. The goal is to make predictions on another set of target-domain unlabeled instances $\{\mathbf{x}_{T_i}\}_{i=1}^{n_T}$. For simplicity, the matrices of source-domain labeled, unlabeled, and all instances are denoted by $\mathbf{X}_{SL} \in R^{l \times d_S}$, $\mathbf{X}_{SU} \in R^{(n_S - l) \times d_S}$ and $\mathbf{X}_{SA} \in R^{n_S \times d_S}$, respectively, where each row corresponds an instance. Similarly, the matrix of target-domain unlabeled instances is denoted by $\mathbf{X}_T \in R^{n_T \times d_T}$, and the corresponding instances in the source and target domains are denoted by $\mathbf{X}_{SC} \in R^{n_C \times d_S}$ and $\mathbf{X}_{TC} \in R^{n_C \times d_T}$ respectively.

As our proposed method is based on matrix completion, we first construct a unified instance-feature matrix for all the instances from the source and target domains as follows,

$$\mathbf{M} = \begin{bmatrix} \overline{\mathbf{X}}_S \\ \overline{\mathbf{X}}_T \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{X}}_{SA} \\ \overline{\mathbf{X}}_{SC} \\ \overline{\mathbf{X}}_T \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{SA} & \mathbf{0}_{n_S, d_T} \\ \mathbf{X}_{SC} & \mathbf{X}_{TC} \\ \mathbf{0}_{n_T, d_S} & \mathbf{X}_T \end{bmatrix}, \qquad (1)$$

where $\overline{\mathbf{X}}_S = \left[\overline{\mathbf{X}}_{SA}^\top \; \overline{\mathbf{X}}_{SC}^\top\right]^\top$, and the matrices $\mathbf{0}_{n_S, d_T}$ and $\mathbf{0}_{n_T, d_S}$ denote the missing correspondences of $\mathbf{X}_{SA}$ and $\mathbf{X}_T$ in the target domain and the source domain, respectively. We further define $n'_S = n_S + n_C$.

We aim to recover the missing entries in $\mathbf{M}$ to obtain the "ground-truth" matrix $\mathbf{X}$. Based on this matrix, we further apply a singular value decomposition (SVD) on $\mathbf{X}$ to project both the source and target domain data to a common latent space spanned by the top $r$ singular vectors, i.e., $\mathbf{X} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r$, and let $\mathbf{Z} = \mathbf{X}\mathbf{V}_r$. Finally we train a classifier on the first $l$ rows of $\mathbf{Z}$, i.e., the new feature representations of the source-domain labeled data, and test on the last $n_T$ rows of $\mathbf{Z}$, i.e., the new feature representations of the target-domain test data.

### Matrix Completion for HTL

Recall that given partially observed cross-domain instances-feature matrix $\mathbf{M}$, we seek to estimate the missing values of the matrices $\mathbf{0}_{n_S, d_T}$ and $\mathbf{0}_{n_T, d_S}$ (i.e., to estimate the missing correspondences of $\mathbf{X}_{SA}$ and $\mathbf{X}_T$ in the target domain and the source domain respectively) by recovering the ground-truth matrix $\mathbf{X}$ that fulfills the following properties:

- $\mathbf{X}$ is sparse. Motivated by a number of applications, such as cross-language text categorization, a feature representation for an instance is sparse. Thus, the $L_1$ norm $\|\cdot\|_1$ can be used to encourage sparsity.
- $\mathbf{X}$ is low rank. In many applications, high dimensional data are controlled by a few latent factors. Thus, the nuclear norm $\|\cdot\|_*$ can be used to regularize its rank.

The HTL problem can then be transformed into a standard matrix completion (MC) problem (Xiao and Guo 2013) as,

$$\min_{\mathbf{X}} \frac{1}{2}\|(\mathbf{X}-\mathbf{M})\circ\mathbf{P}\|_F^2 + \gamma\|\mathbf{X}\|_* + \mu\|\mathbf{X}\|_1 \tag{2}$$

where $\mathbf{M}$ is an observed matrix, $\mathbf{P}$ is an indicator matrix with $\mathbf{P}_{ij} = 1$ if $(i, j)$ is observed, otherwise 0, the operator $\circ$ is the Hadamard product, and $\gamma$, $\mu$ are tradeoff parameters for the nuclear norm $\|\cdot\|_*$ and the $L_1$ norm $\|\cdot\|_1$, respectively.

### Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) (Borgwardt et al. 2006) is a nonparametric measure to estimate the distance between distributions, which is expressed as follows

$$\text{Dist}(\overline{\mathbf{X}}_S, \overline{\mathbf{X}}_T) = \left\| \frac{1}{n'_S}\sum \phi(\overline{\mathbf{x}}_{S_i}) - \frac{1}{n_T}\sum \phi(\overline{\mathbf{x}}_{T_j}) \right\|_{\mathcal{H}}^2, \tag{3}$$

where $\phi(\cdot)$ is a feature mapping to a Hilbert space. It can be rewritten in a more compact form:

$$\text{Dist}(\mathbf{D}_S, \mathbf{D}_T) = \text{tr}(\mathbf{K}_\mathbf{X}\mathbf{L}), \tag{4}$$

where $\mathbf{K}_\mathbf{X} = \begin{pmatrix}\mathbf{K}_{\overline{\mathbf{X}}_{S,S}} & \mathbf{K}_{\overline{\mathbf{X}}_{S,T}} \\ \mathbf{K}_{\overline{\mathbf{X}}_{T,S}} & \mathbf{K}_{\overline{\mathbf{X}}_{T,T}}\end{pmatrix}$, which is a kernel matrix induced by the feature mapping function $\phi(\cdot)$, and $\mathbf{L} = \begin{pmatrix}\mathbf{L}_{S,S} & \mathbf{L}_{S,T} \\ \mathbf{L}_{T,S} & \mathbf{L}_{T,T}\end{pmatrix}$, where $\mathbf{L}_{ij} = 1/(n'_S)^2$ if $\mathbf{x}_i, \mathbf{x}_j \in \overline{\mathbf{X}}_S$, $\mathbf{L}_{ij} = 1/(n_T)^2$ if $\mathbf{x}_i, \mathbf{x}_j \in \overline{\mathbf{X}}_T$, otherwise $\mathbf{L}_{ij} = -1/(n'_S n_T)$.

### Distribution Matching based Matrix Completion

The proposed optimization problem, Distribution Matching based Matrix Completion (DMMC), is expressed as follows:

$$\begin{aligned} \min_{\mathbf{X}\in\mathcal{K}} \quad & F(\mathbf{X}) = \frac{1}{2}\|(\mathbf{X}-\mathbf{M})\circ\mathbf{P}\|_F^2 + \gamma\|\mathbf{X}\|_* + \mu\|\mathbf{X}\|_1 \\ s.t. \quad & \mathcal{K} = \{\mathbf{X}|\text{tr}(\mathbf{K}_\mathbf{X}\mathbf{L}) < \lambda\}, \end{aligned} \tag{5}$$

where $\lambda$ is a constant to constrain the distance between the source and target domain data, $\overline{\mathbf{X}}_S$ and $\overline{\mathbf{X}}_T$, measured by MMD. Note that the difference between (5) and (2) is the additional constraint based on MMD. As shown in experiments, by adding the MMD-based constraint, one can achieve much better performance on HTL problems. This is because the MMD-based regularization term captures the information on distribution difference. In contrast, the rank and sparsity constraints fail to capture such information. For example, a block diagonal matrix satisfies both low rank and sparsity constraints. However, such recovered matrix is not useful for HTL problems since it fails to recover the missing correspondences. Note that obtaining the solution of the optimization problem (5) is nontrivial because the additional MMD-based regularization makes the problem much more difficult to solve. Specifically, by adding the MMD-based regularization, the optimization problem (5) is nonconvex, which cannot be solved by any existing MC algorithms such as singular value thresholding (SVT) (Cai, Candès, and Shen 2010). Therefore, a new approach needs to be developed, which is discussed in the next section.

## HTL through Active Correspondences Construction

In this section, we present the overall framework of Heterogeneous Transfer Learning through Active correspondences construction (HTLA) as described in Algorithm 1. To begin with, we randomly construct several correspondences, which are denoted by a set $\mathcal{C}^{(1)}$, to form the matrix $\mathbf{M}$ in (1), and apply the proposed HTL method DMMC to recover an initial $\mathbf{X}$. After that, we iteratively construct cross-domain correspondences by updating the set $\mathcal{C}^{(t)}$, where $t$ is an index of each iteration, based on the proposed active correspondences construction strategy ActiveLearn$(\cdot)$, and reapply DMMC to recover a more precise $\mathbf{X}$ with the updated $\mathcal{C}^{(t+1)}$ until a stopping criterion is met. In the following sections, we introduce the DMMC method and the strategy of active correspondences construction in detail.

**Algorithm 1** Framework of HTLA

```
**Initializations:** Randomly construct $k$ correspondences $\mathcal{C}^{(1)}$ to form the matrix $\mathbf{M}$, and set $\mathbf{X}^{(1)} = \mathbf{M}$.
**for** $t = 1, 2, ..., T$ **do**
    1: $\mathbf{X}^{(t+1)} = \text{DMMC}(\mathbf{X}^{(t)})$.
    2: Set $\mathcal{C}^{(t+1)} = \text{ActiveLearn}\ (\mathbf{X}^{(t)}, \mathcal{C}^{(t)})$.
    3: Update $\mathbf{X}^{(t+1)}$ based on $\mathcal{C}^{(t+1)}$, and set $t \leftarrow t + 1$.
**end for**
**Return**: $\mathbf{X} \leftarrow \mathbf{X}^{(T+1)}$
**Recover low-dimensional representations:** apply a SVD on $\mathbf{X}$, $\mathbf{X} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r$, and use the top $r$ singular vectors to construct a projection, $\mathbf{Z} = \mathbf{X}\mathbf{V}_r$.
```

### Optimization for DMMC

In this section, we propose an efficient algorithm to solve the optimization problem (5) based on stochastic sub-gradient descent (SSGD) with projection on the feasible space $\mathcal{K}$. The reason that we adopt SSGD is its desirable advantages for large-scale optimization problems (Bottou 2010; Avron et al. 2012). Moreover, its promising results on nonlinear optimization problems have been shown by other researchers (Wang, Crammer, and Vucetic 2012). Specifically, our proposed approach consists of two parts: 1) SSGD for MC, and 2) projection onto $\mathcal{K}$.

**SSGD for MC** The sub-gradient of $F(\mathbf{X})$ in (5) w.r.t $\mathbf{X}$ can be written as

$$\mathcal{G}(F(\mathbf{X})) = \gamma\frac{\partial\|\mathbf{X}\|_*}{\partial\mathbf{X}} + \mu\frac{\partial\|\mathbf{X}\|_1}{\partial\mathbf{X}} + (\mathbf{X}-\mathbf{M})\circ\mathbf{P}, \tag{6}$$

where $\frac{\partial\|\mathbf{X}\|_*}{\partial\mathbf{X}} = \mathbf{U}\mathbf{V}^\top + \mathbf{H}$, and $\mathbf{U\Sigma V}$ is a compact SVD of $\mathbf{X}$.[1]

**Efficient and Unbiased Sub-gradient Estimation:** The success of SSGD relies on the unbiased estimation of the sub-gradient estimation. Avron et al. (2012) proposed to use a probing matrix to sparsify the sub-gradient estimation such that $\mathbb{E}(\tilde{g}) = g$. The probing matrix is defined as follows,

[1] We can simply set $\mathbf{H}$ to be the zero matrix (Watson 1992) and $\frac{\partial\|\mathbf{X}\|_1}{\partial\mathbf{X}_{i,j}} = \text{sign}(\mathbf{X}_{ij})$ when $\mathbf{X}_{ij} \neq 0$, otherwise $\frac{\partial\|\mathbf{X}\|_1}{\partial\mathbf{X}_{i,j}} = \theta \in [-1, 1]$.

**Definition 1. Probing Matrix***: A random $d \times p$ matrix $\mathbf{Y}$ with $p < d$ is a probing matrix if $\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = \mathbf{I}_{d\times d}$, where $I_{d\times d}$ is the identity matrix, and $\mathbb{E}$ is the expectation operator.*

We can randomly sample $p$ vectors from the scaled standard basis $\{\sqrt{d}\mathbf{e}_1, \cdots, \sqrt{d}\mathbf{e}_d\}$ to form the probing matrix: $\mathbf{Y} = [\sqrt{d}\mathbf{e}_1, \cdots, \sqrt{d}\mathbf{e}_p]/\sqrt{p}$. Therefore, we only need to sample $p$ column vectors from the $d$ columns of a $n \times d$ matrix $\mathbf{X}$. As the main time complexity for MC depends on performing a SVD on the matrix, i.e., $O(9d^3+8nd^2+4n^2d)$, where $d = d_S+d_T$ and $n = n'_S+n_T$. Compared to the algorithm used in Xiao and Guo (2013), which performs a SVD on the whole matrix, our proposed method only performs SVD on a submatrix of the original matrix. Therefore, the time complexity is reduced to $O(9p^3 + 8np^2 + 4n^2p)$. Especially, when data are of very high dimensions, $p \ll d$, our proposed approach is much more efficient. Besides, SSGD with unbiased sub-gradient estimation converges in $O(d/p)$ iterations (Theorem 2.3 (Avron et al. 2012)). To avoid additional parameters, we set $p$ of the probing matrix the same as the reduced dimension $r$ as suggested by Avron et al. (2012).

**Projection to $\mathcal{K}$-space** After one step of sub-gradient descent, the recovered $\mathbf{X}$ may be out of the feasible set $\mathcal{K}$. In other words, the distributions of the recovered source- and target- domain data may be very different, which may make cross-domain knowledge transfer unsuccessful. Therefore, the projection on the $\mathcal{K}$-space is required. The operator of $\mathcal{K}$-space projection for any matrix $\mathbf{X}$ is defined as follows

$$\prod_{\mathcal{K}}(\mathbf{X}) = \arg\min_{\hat{\mathbf{X}}\in\mathcal{K}} h(\hat{\mathbf{X}}), \tag{7}$$

where $h(\hat{\mathbf{X}}) = \frac{1}{2}\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2$. The optimization problem (7) is equivalent to the following problem:

$$\min_{\hat{\mathbf{X}}} h(\hat{\mathbf{X}}) = \frac{1}{2}\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda \text{tr}(\mathbf{K}_{\hat{\mathbf{X}}}\mathbf{L}), \tag{8}$$

where we use $\lambda$ as the parameter of the MMD-based term for simplicity in presentation. Though (8) is nonconvex due to some negative terms in $\mathbf{L}$, it can be formulated as the difference of convex program, which can be solved by the concave-convex procedure (CCCP) with global convergence guarantee (Sriperumbudur and Lanckriet 2009) to the minimum or stationary point if the objective function is smooth. To apply CCCP to solve the problem, we rewrite (8) as

$$\begin{aligned}\min_{\hat{\mathbf{X}}} h(\hat{\mathbf{X}}) &= \frac{1}{2}\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda\left(\text{tr}(\mathbf{K}_{\hat{\mathbf{X}}}\mathbf{L}^+) - \text{tr}(\mathbf{K}_{\hat{\mathbf{X}}}\mathbf{L}^-)\right), \\ &= u(\hat{\mathbf{X}}) - v(\hat{\mathbf{X}}), \end{aligned} \tag{9}$$

where $\mathbf{L}^+ = \begin{pmatrix} \mathbf{L}_{S,S} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_{T,T} \end{pmatrix}$, $\mathbf{L}^- = \begin{pmatrix} \mathbf{0} & -\mathbf{L}_{S,T} \\ -\mathbf{L}_{T,S} & \mathbf{0} \end{pmatrix}$, $u(\hat{\mathbf{X}}) = \frac{1}{2}\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda\text{tr}(\mathbf{K}_{\hat{\mathbf{X}}}\mathbf{L}^+)$, and $v(\hat{\mathbf{X}}) = \lambda\text{tr}(\mathbf{K}_{\hat{\mathbf{X}}}\mathbf{L}^-)$.

In general, CCCP solves problems of the following form,

$$\min_{\mathbf{d}} u(\mathbf{d}) - v(\mathbf{d}). \tag{10}$$

If $v(\mathbf{d})$ is smooth, then the updates of $\mathbf{d}$ can be achieved by the majorization-minimization algorithm (Sriperumbudur and Lanckriet 2009) as follows:

$$\mathbf{d}^{(l+1)} \leftarrow \arg\min_{\mathbf{d}} u(\mathbf{d}) - \mathbf{d}^\top \nabla v(\mathbf{d}^{(l)}). \tag{11}$$

However all the existing CCCP solvers are based on the assumption that $\mathbf{d} \in R^d$. Here we extend CCCP in a matrix-form manner as follows,

$$\mathbf{X}^{(l+1)} \leftarrow \arg\min_{\mathbf{X}} u(\mathbf{X}) - \text{tr}(\nabla v(\mathbf{X}^{(l)})^\top \mathbf{X}). \tag{12}$$

By substituting (12) into (9), we obtain

$$\begin{aligned}&\hat{\mathbf{X}}^{(l+1)} \\ &\leftarrow \arg\min_{\hat{\mathbf{X}}} \tfrac{1}{2}\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda\text{tr}(\mathbf{K}_{\hat{\mathbf{X}}}\mathbf{L}^+) - \lambda\text{tr}(\triangledown v(\hat{\mathbf{X}}^{(l)})^\top \hat{\mathbf{X}}).\end{aligned} \tag{13}$$

Note that (13) is convex, whose global optimal solution can be achieved by performing gradient descent approaches. CCCP can always generate the sequence of $\mathbf{X}$ monotonically converging to the stationary point or minimum of the objective function if both $u(\mathbf{X})$ and $v(\mathbf{X})$ are convex and differentiable (Sriperumbudur and Lanckriet 2009). For applications with nonnegative feature values, we can simply use the projected gradient descent algorithm to project the solution into the nonnegative space. The pseudo code of DMMC is presented in Algorithm 2.

**Algorithm 2** SSGD for DMMC

```
Input: Instance-feature matrix $\mathbf{M}$, trade-off parameters $\gamma, \lambda, \mu$, maximum iteration $K$, upper bound of rank $r$.
Initializations: $t = 0$, $\mathbf{X}^{(t)}$ the initial rank-$r$ approximation of $\mathbf{M}$, and stepsize $\eta$.
while $t \le K$ do
    1: Generate an $d \times p$ probing matrix $\mathbf{Y}^{(t)}$
    2: Sub-gradient: $g \leftarrow \mathcal{G}(F(\mathbf{X}^{(t)}))\mathbf{Y}^{(t)}(\mathbf{Y}^{(t)})^\top$
    3: Projection: $\hat{\mathbf{X}}^{(t)} \leftarrow \prod_{\mathcal{K}}(\mathbf{X}^{(t)} - \eta g)$
    4: $\mathbf{X}^{(t+1)} = \hat{\mathbf{X}}^{(t)}$, and $t \leftarrow t + 1$.
end while
Output: $\mathbf{X} = \mathbf{X}^{(K+1)}$
```

## Active Correspondences Construction

In the previous section, we have described a new HTL method, DMMC, where a set of correspondences between domains are given. In this section, we present a strategy to actively select a batch of the most informative instances in the source domain to query their correspondences in the target domain based on DMMC. The motivation is that in matrix completion, some recovered "target-domain instances" based on the corresponding source-domain instances may not be precise, resulting in large distance in distributions between domains. Therefore, such corresponding instances need to be manually constructed. To be specific, the distance in distributions between the recovered source- and target-domain data, $\overline{\mathbf{X}}_S$ and $\overline{\mathbf{X}}_T$, is formulated as follows,

$$\begin{aligned}&\min_{\boldsymbol{\alpha}} \left\| \frac{1}{n'_S - k}\left(\sum_{\mathbf{x}_i\in\overline{\mathbf{X}}_{SLC}} \varphi(\mathbf{x}_i) + \sum_{\mathbf{x}_i\in\overline{\mathbf{X}}_{SU}} \alpha_i\varphi(\mathbf{x}_i)\right) - \frac{1}{n_T}\sum_{\mathbf{x}_i\in\overline{\mathbf{X}}_T}\varphi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2, \\ &s.t.\ \alpha_i \in \{0,1\},\ \boldsymbol{\alpha}^\top\mathbf{1} = n_S - l - k,\end{aligned} \tag{14}$$

where $n'_S = n_S + n_C$, $\overline{\mathbf{X}}_{SLC} = \left[\overline{\mathbf{X}}_{SL}^\top\ \overline{\mathbf{X}}_{SC}^\top\right]^\top$, $k$ is the size of correspondences to be constructed in each iteration, and $\boldsymbol{\alpha}$ is the indicator vector for source-domain unlabeled instances. If the instance $i$ is selected, then $\alpha_i = 1$, which implies that the corresponding recovered target-domain instance does not make the distance in distributions between domains large. Otherwise $\alpha_i = 0$, which implies that the corresponding recovered target-domain instance makes the distributions between domains very different.

Note that (14) is NP-hard due to the integer constraints. We thus propose to further relax the problem into the following convex problem (Chattopadhyay et al. 2013; Gong, Grauman, and Sha 2013):

$$\min_{\boldsymbol{\alpha}\in[0\ 1]} \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{K}_{\overline{\mathbf{X}}_{SU}}\boldsymbol{\alpha} - \frac{n'_S - k}{n_T}\mathbf{k}^\top_{\overline{\mathbf{X}}_{SU},\overline{\mathbf{X}}_T}\boldsymbol{\alpha} + \mathbf{k}^\top_{\overline{\mathbf{X}}_{SU},\overline{\mathbf{X}}_{SLC}}\boldsymbol{\alpha}, \quad (15)$$

where $\mathbf{K}_{\overline{\mathbf{X}}_{SU}}$ is the kernel matrix on $\overline{\mathbf{X}}_{SU}$, $\mathbf{k}_{\overline{\mathbf{X}}_{SU},\overline{\mathbf{X}}_T}(i) = \sum_j \mathbf{K}_{\overline{\mathbf{X}}_{SU},\overline{\mathbf{X}}_T}(i,j)$, and $\mathbf{k}_{\overline{\mathbf{X}}_{SU},\overline{\mathbf{X}}_{SLC}}(i) = \sum_j \mathbf{K}_{\overline{\mathbf{X}}_{SU},\overline{\mathbf{X}}_{SLC}}(i,j)$. After solving (15), we sort the $\{\alpha_i\}$'s in ascending order and select the top $k$ corresponding $\mathbf{x}_{S_i} \in \mathbf{X}_{SU}$ for querying their corresponding instances in the target domain. The overall active correspondences construction procedure is presented in Algorithm 3.

**Algorithm 3** Active Correspondences Construction

```
Input: $\mathbf{X}^{(t)}$, a set of correspondences $\mathcal{C}^{(t)}$ after $t$ iterations, and size $k$ .
1: Compute $\boldsymbol{\alpha}$ for $\overline{\mathbf{x}}_{S_i} \in \overline{\mathbf{X}}_{SU}$ based on (15).
2: Select $\mathcal{C}_S \subseteq \mathbf{X}_{SU}$ of size $k$ whose corresponding $\{\overline{x}_{S_i}\}$'s are of smallest $\{\alpha_i\}$'s values to query their corresponding target-domain instances to form $\mathcal{C}_k$.
3: Update $\mathcal{C}^{(t+1)} = \mathcal{C}^{(t)} \cup \mathcal{C}_k$, and $\mathbf{X}_{SU} = \mathbf{X}_{SU} \setminus \mathcal{C}_S$.
Output: $\mathcal{C}^{(t+1)}$.
```

# Experimental Results

We verify the effectiveness of HTLA by conducting experiments on two cross-language text classification datasets.

## Experimental Setup

**Sentiment Analysis:** The cross-language sentiment classification dataset (Prettenhofer and Stein 2010) comprises of Amazon product reviews of three product categories: books (B), DVDs (D) and music (M). These reviews are written in four languages: English (EN), German (GE), French (FR), and Japanese (JP). For each category, reviews are split into a training set and a test set, including 2,000 reviews respectively. For each non-English language, there are another 2,000 unlabeled correspondences (English v.s. non-English). Each review is preprocessed using TF-IDF.

**Topic Categorization:** The multilingual Reuters collection is a text dataset with up to 5,000 news articles from 6 topics (i.e., C15, CCAT, E21, ECAT, GCAT and M11) in five languages, i.e., English (EN), French (FR), German (GE), Italian (IT) and Spanish (SP), which are represented by a bag-of-words weighted by TF-IDF. Each document has translations in the other four languages. As in practice, English documents are widely accessible, we take English as the source domain, and each of the other languages as a target domain, respectively. The performance of all methods are evaluated on the target-domain unlabeled data without any target-domain labeled training data.

**Experimental Design:** We conduct three sets of experiments to evaluate the performance of our proposed framework for HTL. The first experiment compares DMMC with other HTL methods when a set of cross-domain correspondences are given. The second experiment compares the performance of our proposed active strategy with random selection for cross-domain correspondences construction. The third experiment studies the sensitivity of the parameters in our proposed framework.

**Baselines:** We compared our proposed DMMC for HTL with the following state-of-the-art baselines.

1) **SVM-SC:** We first train a classifier on the source-domain labeled data, and then use it to make predictions on the source-domain corresponding data. In this way, the predicted labels on the source-domain corresponding data can be transferred to their correspondences, i.e., translations, in the target domain .

2) **OPCA:** We first apply Oriented Principal Component Analysis (OPCA) (Platt, Toutanova, and Yih 2010) to learn projections for the source and target domain data. A classifier is trained and tested on the projected data.

3) **LSI:** We first apply Latent Semantic Indexing (LSI) (Nie et al. 1999) on the matrix $\mathbf{M}$ to learn low-dimensional cross-lingual representations, and then train and test a classifier on the low-dimensional data.

4) **KCCA:** We first apply Kernel Canonical Component Analysis (KCCA) (Vinokourov, Shawe-Taylor, and Cristianini 2002) on the cross-domain correspondences to learn projections for bilingual data, and then train and test a classifier on the projected data.

5) **TSL:** We apply the correspondence-based HTL method proposed by Xiao and Guo (2013) to train a classifier to predict the unlabeled data in the target domain.

**Parameter Settings:** For DMMC, there are three tradeoff parameters, $\gamma$, $\mu$, and $\lambda$, on the regularization terms on low-rank, sparsity, and distance matching, respectively. We validate the parameters as follows: $\gamma \in \{10^{-2}, 10^{-1}, ..., 10^{2}\}$, $\mu \in \{10^{-6}, 10^{-5}, ..., 10^{-1}\}$, and $\lambda \in \{10^{-3}, 10^{-2}, ..., 10^{1}\}$. For the parameter setting of TSL, we follow the procedure in (Xiao and Guo 2013). We first set $\mu = 10^{-6}$ and $\tau = 1$, and then cross validate $\gamma \in \{10^{-3}, 10^{-2}, ..., 10^{1}\}$, and $\rho \in \{10^{-6}, 10^{-5}, ..., 10^{-1}\}$. For KCCA, we tune the parameter $\kappa$ in (5) in (Vinokourov, Shawe-Taylor, and Cristianini 2002) in the range of $\{10^{-2}, 10^{-1}, ..., 10^{2}\}$. For the reduced dimensions for OPCA, LSI, KCCA, TSL and DMMC, we validate in the same range of $r \in \{20, 50, 100, 200, 500\}$. For all experiments, we employ linear support vector machines (SVMs) (Chang and Lin 2011) with default parameter settings as the base classifier.

## Overall Comparison Results

In the first experiment, we compare the performance in terms of classification accuracy of different methods given all the available cross-domain correspondences in the datasets. The overall results are shown in Table 1 and 2. We can see that DMMC outperforms all the baselines. SVM-SC performs

Table 1: Sentiment analysis in classification accuracy (%).

| TASK | SVM-SC | OPCA | KCCA | LSI | TSL | DMMC |
|---|---|---|---|---|---|---|
| B-FR | 72.35 | 68.70 | 69.75 | 71.56 | 73.95 | **76.52** |
| D-FR | 74.50 | 65.45 | 72.55 | 71.22 | 74.30 | **76.23** |
| M-FR | 67.40 | 67.95 | 70.17 | 67.39 | 71.15 | **74.05** |
| B-GE | 74.12 | 69.00 | 73.69 | 72.43 | 75.98 | **77.47** |
| D-GE | 73.75 | 71.63 | 74.79 | 74.76 | 76.01 | **78.28** |
| M-GE | 73.25 | 65.12 | 69.86 | 72.18 | 74.57 | **76.60** |
| B-JP | 60.83 | 58.69 | 62.02 | 62.22 | 65.81 | **68.54** |
| D-JP | 69.36 | 66.60 | 66.17 | 66.47 | 70.72 | **72.12** |
| M-JP | 65.59 | 64.40 | 63.80 | 65.54 | 68.22 | **71.37** |

Table 2: Topic categorization in classification accuracy (%).

| TASK | SVM-SC | OPCA | KCCA | LSI | TSL | DMMC |
|---|---|---|---|---|---|---|
| FR | 63.77 | 58.70 | 63.41 | 61.77 | 63.18 | **65.52** |
| GE | 47.59 | 53.18 | 55.71 | 54.48 | 56.08 | **58.23** |
| IT | 49.23 | 52.95 | 56.17 | 55.39 | 57.15 | **60.76** |
| SP | 53.63 | 55.00 | 56.69 | 55.53 | 56.98 | **62.64** |

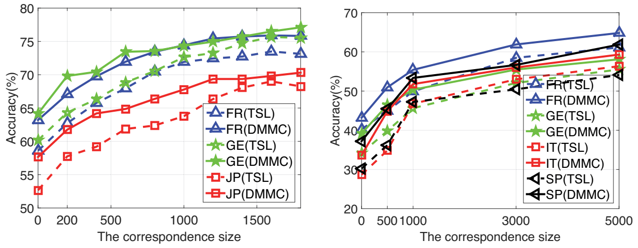
slightly better than OPCA, LSI and KCCA on average due to the sufficient set of cross-domain correspondences. LSI performs poorly without filling the missing values via matrix completion. Therefore, matrix completion is necessary before dimensionality reduction. DMMC explicitly minimizes the distance in distributions of the recovered data between domains. This improves classification accuracy by around 3% over TSL with all the available correspondences. To further analyze the effect of the MMD-based regularization term, we show how the accuracy varies with different correspondence sizes for TSL and DMMC in Figure 2. DMMC outperforms TSL significantly by around 5% in terms of accuracy, especially when the correspondence size is small.

## Experiments on Active Correspondences

In the second experiment, we aim to verify the effectiveness of the proposed active learning strategy for cross-domain correspondences construction. We denote by DMMC-active and DMMC-rand the DMMC method with active and random correspondences construction strategies, respectively. We set $k = 100$, and show the comparison results in Figure 3, where averaged results of each language over 10 runs are reported. We observe that DMMC-active consistently and significantly outperforms DMMC-rand on all the tasks especially when the correspondence size is small. Moreover, DMMC-active performs more stable than DMMC-rand because DMMC-active aims to choose the most informative correspondences to update the model in each iteration.

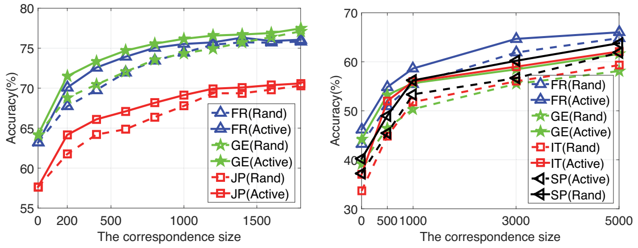
## Parameter Analysis

For the final experiment, we study the sensitivity of the three parameters, $\gamma$, $\mu$, and $\lambda$, in DMMC. In this experiment[2], we fix the correspondences size $n_c = 2,000$ and report averaged results of each non-English language as the target domain in Figure 4. In Figure 4(a), we show the classification accuracy when varies $\gamma$ in the range $[10^{-2}\ 10^{2}]$ by fixing $\mu = 10^{-4}$ and $\lambda = 10^{-2}$. From the figure, we observe that classification accuracy drops fast when $\gamma$ becomes small ($\leq 1$). This implies that the low-rank regularization is important because documents are represented by only a few latent topics. In Figure 4(b), we vary $\mu$ in the range $[10^{-6}\ 10^{-1}]$ by fixing $\gamma = 10$, and $\lambda = 10^{-2}$ . We see that if $\mu$ is set to a large value, then the model may overfit to the observed nonzero entries, and lead to many recovered entries to be zero values. In Figure 4(c), we show classification accuracy under different $\lambda$ in the range $[10^{-3}\ 10^{1}]$ by fixing $\gamma = 10$, and $\mu = 10^{-4}$. The figure shows that the parameter $\lambda$ should be neither too large nor too small to reach satisfactory performance.



(a) Sentiment Analysis (b) Topic Categorization

Figure 2: TSL v.s. DMMC.



(a) Sentiment Analysis (b) Topic Categorization

Figure 3: Active Correspondences Construction.

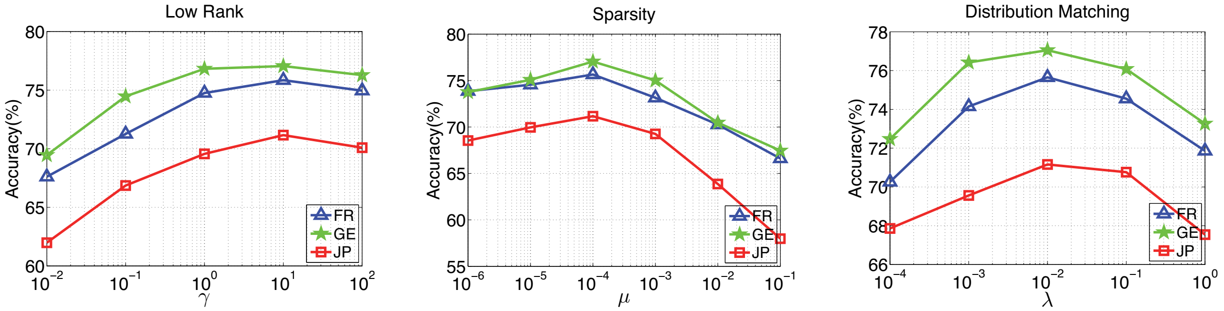
## Conclusion

In this paper, we propose a general framework named HTL through active correspondences construction (HTLA). Different from previous work that assumed a sufficient set of cross-domain correspondences be given in advance, we propose to actively construct cross-domain correspondences. Under the framework, we first propose a general correspondence-based HTL method through matrix completion with a distribution matching regularizer. Based on this method, we propose an active learning strategy to query construction of correspondences. Extensive experiments on two benchmark datasets on multi-language text classification demonstrate the superiority of the proposed method over a number of state-of-the-art baseline methods.

## Acknowledgments

This work is supported by the NTU Singapore Nanyang Assistant Professorship (NAP) grant M4081532.020, the ARC Future Fellowship FT130100746, ARC grant LP150100671, and Singapore MOE AcRF Tier-1 grant RG41/12.

[2] We use sentiment dataset as a showcase due to page limit.



(a) Low Rank (b) Sparsity (c) Distribution Matching

Figure 4: Parameter Analysis.

## References

Avron, H.; Kale, S.; Kasiviswanathan, S. P.; and Sindhwani, V. 2012. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In *ICML*.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22(14):e49–e57.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*. 177–186.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27:1–27:27.

Chattopadhyay, R.; Fan, W.; Davidson, I.; Panchanathan, S.; and Ye, J. 2013. Joint transfer and batch-mode active learning. In *ICML*, 253–261.

Dai, W.; Chen, Y.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2008. Translated learning: Transfer learning across different feature spaces. In *NIPS*, 353–360.

Daumé III, H. 2007. Frustratingly easy domain adaptation. In *ACL*, 256–263.

Duan, L.; Xu, D.; and Tsang, I. W. 2012. Learning with augmented features for heterogeneous domain adaptation. In *ICML*.

Gong, B.; Grauman, K.; and Sha, F. 2013. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 222–230.

Kulis, B.; Saenko, K.; and Darrell, T. 2011. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 1785–1792.

Nie, J.-Y.; Simard, M.; Isabelle, P.; and Durand, R. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR*, 74–81.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22(10):1345–1359.

Platt, J. C.; Toutanova, K.; and Yih, W.-t. 2010. Translingual document representations from discriminative projections. In *EMNLP*, 251–261.

Prettenhofer, P., and Stein, B. 2010. Cross-language text classification using structural correspondence learning. In *ACL*, 1118–1127.

Rai, P.; Saha, A.; Daumé III, H.; and Venkatasubramanian, S. 2010. Domain adaptation meets active learning. In *NAACL HLT 2010 Workshop on ALNLP*, 27–32.

Settles, B. 2010. Active learning literature survey.

Shi, X.; Fan, W.; and Ren, J. 2008. Actively transfer domain knowledge. In *ECML/PKDD*. 342–357.

Sriperumbudur, B. K., and Lanckriet, G. R. G. 2009. On the convergence of the concave-convex procedure. In *NIPS*, 1759–1767.

Vinokourov, A.; Shawe-Taylor, J.; and Cristianini, N. 2002. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS*, 1473–1480.

Wang, C., and Mahadevan, S. 2011. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, 1541–1546.

Wang, Z.; Crammer, K.; and Vucetic, S. 2012. Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale svm training. *J. Mach. Learn. Res.* 13(1):3103–3131.

Wang, X.; Huang, T.-K.; and Schneider, J. 2014. Active transfer learning under model shift. In *ICML*, 1305–1313.

Watson, G. A. 1992. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications* 170:33–45.

Xiao, M., and Guo, Y. 2013. A novel two-step method for cross language representation learning. In *NIPS*, 1259–1267.

Zhao, L.; Pan, S. J.; Xiang, E. W.; Zhong, E.; Lu, Z.; and Yang, Q. 2013. Active transfer learning for cross-system recommendation. In *AAAI*.

Zhou, J. T.; Pan, S. J.; Tsang, I. W.; and Yan, Y. 2014a. Hybrid heterogeneous transfer learning through deep learning. In *AAAI*, 2213–2220.

Zhou, J. T.; Tsang, I. W.; Pan, S. J.; and Tan, M. 2014b. Heterogeneous domain adaptation for multiple classes. In *AISTATS*, 1095–1103.