Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Topic Correlation Analysis for Cross-Domain Text Classification

Lianghao Li[†] **Xiaoming Jin**[†] **Mingsheng Long**[‡]
[†]Key Laboratory for Information System Security, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology
School of Software, Tsinghua University, Beijing, China
[‡]Department of Computer Science and Technology, Tsinghua University, Beijing, China
lianghaoli@yahoo.com xmjin@tsinghua.edu.cn longmingsheng@gmail.com

Abstract

Cross-domain text classification aims to automatically train a precise text classifier for a target domain by using labeled text data from a related source domain. To this end, the distribution gap between different domains has to be reduced. In previous works, a certain number of shared latent features (e.g., latent topics, principal components, etc.) are extracted to represent documents from different domains, and thus reduce the distribution gap. However, only relying the shared latent features as the domain bridge may limit the amount of knowledge transferred. This limitation is more serious when the distribution gap is so large that only a small number of latent features can be shared between domains. In this paper, we propose a novel approach named Topic Correlation Analysis (TCA), which extracts both the shared and the domain-specific latent features to facilitate effective knowledge transfer. In TCA, all word features are first grouped into the shared and the domain-specific topics using a joint mixture model. Then the correlations between the two kinds of topics are inferred and used to induce a mapping between the domain-specific topics from different domains. Finally, both the shared and the mapped domain-specific topics are utilized to span a new shared feature space where the supervised knowledge can be effectively transferred. The experimental results on two real-world data sets justify the superiority of the proposed method over the stat-of-the-art baselines.

Introduction

Text classification algorithms have been proven to be effective in automatically organizing text data. In practice, however, it is usually expensive to obtain sufficient labeled documents to train a precise classifier for a certain domain, whereas there are plenty of labeled documents in a related but different domain. So it would be favorable if we can leverage the labeled documents from the related domain to train a precise classifier for the target domain. However, since different domains usually differ in their underlying distributions, the traditional classification algorithms would be challenged when training data and test data come from different domains (Dai et al. 2007).

Recently, cross-domain text classification has been proposed to solve the above problem. A key idea of many previous works (Xue et al. 2008; Gupta et al. 2010; Pan et al. 2010) is that even if the two domains are different in data distributions, there may exist some latent features (e.g., latent topics, principal components, etc.) that do not cause changes between the domains. If we use these shared latent features to represent documents in a new feature space, we can transfer the supervised knowledge between domains in that space. However, in many real-world applications, it is hard to identify a reasonable number of latent features that are shared by different domains, especially when the distribution gap is large. Therefore, how to further extract the domain-specific latent features for knowledge transfer is crucial to improve the performance of cross-domain text classification.

To address this problem, in this paper, we propose a novel approach named Topic Correlation Analysis (TCA) for cross-domain text classification. TCA extracts both the shared and the domain-specific latent features as a bridge to transfer the text classification knowledge between domains. More specifically, in TCA, we first jointly model the documents from different domains with two kinds of topics: one is shared by different domains and the other is specific to each domain. Then the shared topics are leveraged to identify the correlations between domain-specific topics from different domains. The key idea of our method is that if two domain-specific topics are related to many shared topics in the similar manner, then they tend to be semantically correlated and should be mapped to each other. Finally, by using both the shared and the mapped domain-specific topics to represent documents, we can construct a new shared space spanned by the two kinds of topics, where the distribution gap is greatly reduced. Ideally, if we can correctly induce a mapping between the domain-specific topics from different domains, more supervised knowledge can be transferred through them to achieve improved performance for cross-domain text classification.

To evaluate the effectiveness of the proposed method, we conducted experiments on two real-world text data sets. The experimental results demonstrate that our method can effectively identify the correlations between domain-specific topics, and improve the classification accuracy compared with the state-of-the-art baselines.

Related Work

Cross-domain text classification is a new learning setting which allows training and test documents to come from different distributions (Dai et al. 2007). Existing approaches in this direction can be generally put into two categories: feature-representation based methods (Blitzer, McDonald, and Pereira 2006; Dai et al. 2007; Xue et al. 2008) and instance-weighting based approaches (Jiang and Zhai 2007).

The main idea of feature-representation based methods is to discover a latent feature space where the distributions of different domains are drawn closer. To address this problem, Blitzer et al. proposed the Structural Correspondence Learning (SCL) (Blitzer, McDonald, and Pereira 2006) algorithm which develops a shared latent space by modeling the relationship between the "pivot" and the "non-pivot" features among a set of pre-defined tasks. Pan et al. proposed the Spectral Feature Alignment (SFA) (Pan et al. 2010) algorithm for cross-domain sentiment classification. In SFA, a set of domain-independent sentiment words are identified at first. Then the spectral clustering algorithm is adapted to co-cluster the "domain-independent" and the "domain-specific" features into a shared cluster space. Unlike SCL and SFA which aim to model the relationship between word features, we propose to group word features into high-level semantic topics and then find the correlations between these topics to bridge domains. This leads to a more compact representation of documents for knowledge transfer. Besides, their heuristic selection criteria of the shared features (i.e., the "pivot" features in SCL and the domain-independent sentiment words in SFA) may be sensitive to different applications, whereas our method aims to extract the shared topics with a well-defined probabilistic model.

Recently, topic modeling methods have been adapted for cross-domain text classification. Xue et al. extended PLSA to jointly model the documents from different domains (Xue et al. 2008). However, in their method all topics are assumed to be shared by different domains, which may be too strict in practice. Zhuang et al. proposed the Collaborative Dual-PLSA (CDPLSA) (Zhuang et al. 2010) method to decompose the words and the documents from both domains into word clusters (i.e., topics) and documents clusters (i.e., document categories). In CDPLSA, the associations between topics and document categories are assumed to be stable across domains. Gupta et al. proposed to learn a shared subspace by jointly extracting the common and the domain-specific bases with the Joint Shared Nonnegative Matrix Factorization (JSNMF) (Gupta et al. 2010). The learned shared subspace is used to bridge different domains. Different from them, we explicitly extract the shared and the domain-specific features, and utilize the correlations between them for knowledge transfer.

Modeling documents from different collections has also been studied in text mining. Zhai et al. proposed a cross-collection mixture model, in which each concept is modeled with one common topic and many collection-specific topics (Zhai, Velivelli, and Yu 2004). In their model, each collection-specific topic must be strictly correlated to one common topic. However, this assumption may be too strong in the cross-domain classification, since different domains may have some domain-specific topics which are not correlated to any common topics.

Problem Definition and Preliminaries

In this section, we first define the problem to be addressed. Then we briefly review the Probabilistic Latent Semantic Analysis (PLSA) which is a building block of our method.

Definition 1 *(Cross-Domain Text Classification) Given a source domain $\mathcal{D}^s = \{(d_1^s, y_1^s), \ldots, (d_{N^s}^s, y_{N^s}^s)\}$ consists of N^s labeled documents, and a target domain $\mathcal{D}^t = \{d_1^t, \ldots, d_{N^t}^t\}$ consists of N^t unlabeled documents. Let $\mathcal{W}$ be the vocabulary and $\mathcal{Y}$ be the pre-defined label set. The task is to train a precise classifier $f^t : \mathcal{D}^t \rightarrow \mathcal{Y}$ for predicting the class labels of unlabeled documents in the target domain.*

For example, a set of labeled newsgroup documents can be regarded as a source domain, and a set of unlabeled posters from personal blogs can be regarded as a target domain. The task is to utilize the labeled newsgroup documents to build a precise classifier for predicting the class labels of unlabeled posters.

A Brief Review of PLSA Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999) has been widely used for topic modeling. It assumes the following generative process for word/document co-occurrences:

- select a document d_i with probability $P(D = d_i)$,
- draw a topic z_k with probability $P(Z = z_k | D = d_i)$,
- select a word w_j with probability $P(W = w_j | Z = z_k)$.

The probabilities of $P(D = d_i)$, $P(Z = z_k | D = d_i)$ and $P(W = w_j | Z = z_k)$ over $\{d_i, z_k, w_j\}_{i,j,k}$ are estimated by maximizing the likelihood of all observed word/document co-occurrences.

Our Solution

This section presents our proposed Topic Correlation Analysis (TCA) method for cross-domain text classification. We first present how to extract both the shared and the domain-specific topics by jointly modeling text data in both domains. Then we discuss how to use the shared topics to induce a mapping between the domain-specific topics from different domains. Finally, we describe how to utilize the shared and the mapped domain-specific topics to construct a new shared space for cross-domain text classification. The frequently used notations are listed in Table 1.

Mining Shared and Domain-Specific Topics

In this part, we propose a novel Joint Mixture Model (JMM) which adapts PLSA to model both the shared and the domain-specific topics. In JMM, the co-occurrence of a document and a word is associated with a latent topic $Z \in \mathcal{Z} = \{z_1, \ldots, z_K, z_1^s, \ldots, z_{K^s}^s, z_1^t, \ldots, z_{K^t}^t\}$ which is drawn from either the shared topics, i.e. $\{z_1, \ldots, z_K\}$, or the domain-specific topics, i.e. $\{z_1^s, \ldots, z_{K^s}^s, z_1^t, \ldots, z_{K^t}^t\}$. A latent decision variable $\pi \in \{0, 1\}$ is introduced to control which kind of topics is drawn. The generative process for the word/document co-occurrence in JMM is defined as:

Table 1: Notations

Symbols	Description
ℓ	domain random variable, $\ell \in \{s, t\}$
D^ℓ	document random variable, $D^\ell \in \mathcal{D}^\ell$
Z	topic random variable, $Z \in \mathcal{Z}$
π	decision random variable, $\pi \in \{0, 1\}$
W	word random variable, $W \in \mathcal{W}$
d_n^ℓ	nth document in domain ℓ
w_j	jth word in the vocabulary
z_k	kth shared topic
z_r^ℓ	rth domain-specific topic in domain ℓ
$\mu_{d_i^\ell}$	document-level parameter for the decision process
K	number of the shared topics
K^ℓ	number of the domain-specific topics in domain ℓ
$\mathcal{W}$	vocabulary

- select a document d_i^ℓ with probability $P(D^\ell = d_i^\ell)$,
- draw a decision random variable $\pi \sim \text{Bern}(\mu_{d_i^\ell})$,

1. if $\pi = 0$, pick a shared topic z_k with probability $P(Z = z_k|D^\ell = d_i^\ell, \pi = 0)$, and generate a word w_j with probability $P(W = w_j|Z = z_k)$,
2. else if $\pi = 1$, pick a domain-specific topic z_r^ℓ with probability $P(Z = z_r^\ell|D^\ell = d_i^\ell, \pi = 1)$, and generate a word w_j with probability $P(W = w_j|Z = z_r^\ell)$.

Here $\mu_{d_i^\ell}$ is a document-level parameter for the decision process. After integrating out the latent variables Z and π, one can obtain the observation pair (d_i^ℓ, w_j) as a result. The joint probability model for the data generative process described above is[1]:

$$\begin{aligned}
P(d_i^\ell, w_j) &= P(d_i^\ell)P(w_j|d_i^\ell), \quad \text{where} \\
P(w_j|d_i^\ell) &= P(\pi = 0|\mu_{d_i^\ell}) \sum_{k=1}^{K} P(z_k|d_i^\ell, \pi = 0)P(w_j|z_k) \\
&\quad + P(\pi = 1|\mu_{d_i^\ell}) \sum_{r=1}^{K^\ell} P(z_r^\ell|d_i^\ell, \pi = 1)P(w_j|z_r^\ell).
\end{aligned} \tag{1}$$

Taking the product of the probabilities of single documents, we obtain the log-likelihood of all documents:

$$\mathcal{L} = \sum_{\ell \in \{s,t\}} \sum_{d_i^\ell \in \mathcal{D}^\ell} \sum_{w_j \in \mathcal{W}} n(d_i^\ell, w_j) \log P(d_i^\ell, w_j), \tag{2}$$

where $n(d_i^\ell, w_j)$ denotes the number of times that word w_j occurs in document d_i^ℓ.

Without any prior knowledge, we adopt the maximum likelihood estimators (MLE) to estimate the parameters. A general way to find the maximum likelihood solution for latent variable models is the Expectation Maximization (EM) algorithm (Bishop 2006). The EM algorithm iteratively optimizes the likelihood function with two steps: an expectation (E) step, where the posterior probabilities for latent variables are evaluated with the current values of parameters; and a maximization (M) step, where the parameters are updated by maximizing the expected completed likelihood which depends on the evaluated posterior probabilities in the E-step. After incorporating the normalization constraints, one can obtain the E-step and the M-step updates as presented in Figure 1.

In order to leverage the shared topics to find the correlations between domain-specific topics, we need to obtain sufficient co-occurrences between the two kinds of topics. To this end, we add a symmetric Beta distribution Beta$(\alpha + 1)$ as the prior distribution for $\{\mu_i^\ell\}_{\ell,i}$ to smooth the decision process. Since the Beta distribution is the conjugate prior of the Bernoulli distribution, the symmetric hyperparameter α can be interpreted as pseudo-observations of the shared and the domain-specific topics in each document. The larger α is, the more likely a document be equally generated by the two kinds of topics. In this case, we adopt the EM algorithm to find the maximum posterior (MAP) solution for the parameters. In our model, the E-step remains the same as that in MLE, whereas in the M-step the update for μ_i^ℓ should be replaced with

$$\mu_i^\ell = \frac{\sum_{w_j} n(d_i^\ell, w_j) \sum_{k=1}^{K} P(z_k, \pi = 0|d_i^\ell, w_j) + \alpha}{\sum_{w_j} n(d_i^\ell, w_j) + 2 \cdot \alpha} \tag{3}$$

The M-step updates for other parameters are unchanged.

Measuring Topic Correlations across Domains

In this part, we discuss how to measure the correlations between domain-specific topics from different domains.

Since there is no directly co-occurrences between the domain-specific topics across domains, we propose to leverage the shared topics as a bridge to find the correlations between them. Our observation is that if two domain-specific topics are related to many shared topics in the same way, they tend to be semantically correlated. So we need to: 1) calculate the similarity between the shared and the domain-specific topics in each domain, and 2) infer the correlations between the domain-specific topics from different domains.

For the first subproblem, we adopt the Jensen-Shannon divergence between the document-topic distributions (i.e., $P(D^\ell|Z)$) as the similarity measure. The Jensen-Shannon divergence is widely used for measuring the similarity between two probability distributions (Lin 1991). It is a symmetrized and smoothed version of the Kullback-Leibler divergence. Specifically, for two distributions P and Q, the Jensen-Shannon divergence between them is:

$$\text{JSD}\,(P \parallel Q) = \frac{1}{2}\text{KL}(P \parallel M) + \frac{1}{2}\text{KL}(Q \parallel M),$$

where $M = \frac{1}{2}(P + Q)$ and KL$(\cdot\|\cdot)$ is the Kullback-Leibler divergence between the two distributions. So the similarity $\theta_{z_r^\ell, z_k}$ between shared topic z_k and domain-specific topic z_r^ℓ is given by:

$$\theta_{z_r^\ell, z_k} = \text{JSD}\left(P(D^\ell|Z = z_k) \parallel P(D^\ell|Z = z_r^\ell)\right). \tag{4}$$

[1] For simplicity, we omit the names of random variables when this causes no confusions.

E-step:

$$P(z_k, \pi = 0 | d_i^\ell, w_j) = \frac{\mu_{d_i^\ell} P(z_k | d_i^\ell, \pi = 0) P(w_j | z_k)}{\mu_{d_i^\ell} \sum_{k=1}^{K} P(z_k | d_i^\ell, \pi = 0) P(w_j | z_k) + (1 - \mu_{d_i^\ell}) \sum_{r=1}^{K^\ell} P(z_r^\ell | d_i^\ell, \pi = 1) P(w_j | z_r^\ell)}$$

$$P(z_r^\ell, \pi = 1 | d_i^\ell, w_j) = \frac{(1 - \mu_{d_i^\ell}) P(z_r^\ell | d_i^\ell, \pi = 1) P(w_j | z_r^\ell)}{\mu_{d_i^\ell} \sum_{k=1}^{K} P(z_k | d_i^\ell, \pi = 0) P(w_j | z_k) + (1 - \mu_{d_i^\ell}) \sum_{r=1}^{K^\ell} P(z_r^\ell | d_i^\ell, \pi = 1) P(w_j | z_r^\ell)}$$

M-step:

$$P(w_j | z_k) = \frac{\sum_\ell \sum_{d_i^\ell} n(d_i^\ell, w_j) P(z_k, \pi = 0 | d_i^\ell, w_j)}{\sum_{w_n} \sum_\ell \sum_{d_i^\ell} n(d_i^\ell, w_n) P(z_k, \pi = 0 | d_i^\ell, w_n)}, \quad P(z_k | d_i^\ell, \pi = 0) = \frac{\sum_{w_j} n(d_i^\ell, w_j) P(z_k, \pi = 0 | d_i^\ell, w_j)}{\sum_{m=1}^{K} \sum_{w_j} n(d_i^\ell, w_j) P(z_m, \pi = 0 | d_i^\ell, w_j)}$$

$$P(w_j | z_r^\ell) = \frac{\sum_{d_i^\ell} n(d_i^\ell, w_j) P(z_r^\ell, \pi = 1 | d_i^\ell, w_j)}{\sum_{w_n} \sum_{d_i^\ell} n(d_i^\ell, w_n) P(z_r^\ell, \pi = 1 | d_i^\ell, w_n)}, \quad P(z_r^\ell | d_i^\ell, \pi = 1) = \frac{\sum_{w_j} n(d_i^\ell, w_j) P(z_r^\ell, \pi = 1 | d_i^\ell, w_j)}{\sum_{m=1}^{K^\ell} \sum_{w_j} n(d_i^\ell, w_j) P(z_m^\ell, \pi = 1 | d_i^\ell, w_j)}$$

$$\mu_{d_i^\ell} = \frac{\sum_{w_j} n(d_i^\ell, w_j) \sum_{k=1}^{K} P(z_k, \pi = 0 | d_i^\ell, w_j)}{\sum_{w_j} n(d_i^\ell, w_j)}.$$

Figure 1: EM updates for the Joint Mixture Model

An intuitive explanation of the above measure is that if a shared topic and a domain-specific topic always co-occur with each other in the documents, they should have low Jensen-Shannon divergence and tend to be related.

For the second subproblem, since the domain-specific topics never co-occurred across domains, we cannot directly evaluate their correlations using (4). In this paper, we leverage the similarity between the shared and the domain-specific topics to infer the correlations between them. Specifically, the Pearson's Correlation Coefficients (PCCs) is adopted to calculate their correlations. In statistics, PCCs is used to calculate the correlation between two random variables. Recently, it has been adopted to measure the user/item similarity in the memory-based methods for Collaborative Filtering (Breese, Heckerman, and Kadie 1998). With PCCs, the correlation $\rho(z_i^t, z_j^s)$ between domain-specific topic z_j^s in the source domain and domain-specific topic z_i^t in the target domain is:

$$\rho(z_i^t, z_j^s) = \frac{\sum_k (\theta_{z_i^t, z_k} - \overline{\theta}_{z_i^t})(\theta_{z_j^s, z_k} - \overline{\theta}_{z_j^s})}{\sqrt{\sum_l (\theta_{z_i^t, z_l} - \overline{\theta}_{z_i^t})^2} \sqrt{\sum_l (\theta_{z_j^s, z_l} - \overline{\theta}_{z_j^s})^2}}, \tag{5}$$

where $\overline{\theta}_{z_i^\ell}$ is the mean similarity between z_i^ℓ and all the shared topics. According to the property of PCCs, a positive value of ρ indicates that two domain-specific topics are simultaneously related, or simultaneously unrelated, to the shared topics. And a negative value of ρ indicates that when the relatedness between one domain-specific topic and the shared topics increases, the relatedness between topics for the other will decrease. With the learned topic correlations, a topic mapping matrix $\mathbf{U} \in \mathbb{R}^{K^t \times K^s}$ can be constructed as follows:

$$\mathbf{U} = \begin{bmatrix} \rho(z_1^t, z_1^s) & \rho(z_1^t, z_2^s) & \cdots & \rho(z_1^t, z_{K^s}^s) \\ \rho(z_2^t, z_1^s) & \rho(z_2^t, z_2^s) & \cdots & \rho(z_2^t, z_{K^s}^s) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(z_{K^t}^t, z_1^s) & \rho(z_{K^t}^t, z_2^s) & \cdots & \rho(z_{K^t}^t, z_{K^s}^s) \end{bmatrix} \tag{6}$$

Constructing New Representation for Documents

Now we describe how to map documents from both domains into a new shared space for cross-domain text classification. First, we represent documents $\{d_i^\ell\}_{\ell,i}$ from both domains with the extracted topics:

$$\phi(d_i^\ell) = \left[P(z_1 | d_i^\ell), \ldots, P(z_K | d_i^\ell), P(z_1^\ell | d_i^\ell), \ldots, P(z_{K^\ell}^\ell | d_i^\ell) \right]^{\mathrm{T}} \tag{7}$$

Then we transform the domain-specific part representation $\phi(d_i^t)_{[K+1:K+K^t]}$ for documents in the target domain with the topic mapping matrix $\mathbf{U}$. The new representation $\psi(\phi(d_i^t))$ of document d_i^t from the target domain is given by:

$$\psi(\phi(d_i^t)) = \left[\phi(d_i^t)_{[1:K]}; \mathbf{U}^{\mathrm{T}} \phi(d_i^t)_{[K+1:K+K^t]} \right], \tag{8}$$

which consists of two parts: one is represented by the shared topics, and the other is represented by the mapped source domain-specific topics. With the new representation, each document in the target domain is mapped into the new feature space spanned by the shared topics and the source domain-specific topics. Now one can train a conventional classifier in the new feature space by using the labeled documents $\mathcal{D}^s = \{(\phi(d_1^s), y_1^s), \ldots, (\phi(d_{N^s}^s), y_{N^s}^s)\}$ from the source domain, and predict the class labels of the unlabeled documents $\mathcal{D}^t = \{\psi(\phi(d_1^t)), \ldots, \psi(\phi(d_{N^t}^t))\}$ in the target domain.

The complete process of our proposed method is summarized in Algorithm 1.

Experiments

In this section, we conduct experiments on two real-world data sets to verify the effectiveness of our proposed Topic Correlation Analysis (TCA) method for cross-domain text classification.

Data Sets

20Newsgroups The 20Newsgroups data set has been widely used for evaluating the performance of cross-domain

Algorithm 1: Cross-Domain Text Classification via Topic Correlation Analysis

```
Input  : (1) Source domain labeled data set $\mathcal{D}^s$ and
         target domain unlabeled data set $\mathcal{D}^t$, (2)
         Number of shared topics $K$, (3) Number of
         source/target domain-specific topics $K^s$, $K^t$,
         (4) Number of iterations $T$
Output: The predicted class label of each unlabeled
         document $d_i^t \in \mathcal{D}^t$ in the target domain
Initialize the parameters of the proposed JMM model;
for $t \leftarrow 1$ to $T$ do
    E-step: Compute the posterior probabilities with
    the E-step updates in Figure 1;
    M-step: Update the model parameters with the
    M-step updates in Figure 1 and (3);
end
Measure the topic correlations using (4) and (5);
Construct the topic mapping matrix $\mathbf{U}$ using (6);
Represent each document with extracted topics as (7);
Transform each document in the target domain as (8);
Train a classifier with the labeled documents
$\mathcal{D}^s = \{(\phi(d_n^s), y_n^s)\}_{n=1}^{N^s}$ and predict the class labels of
the unlabeled documents $\mathcal{D}^t = \{\psi(\phi(d_i^t))\}_{i=1}^{N^t}$.
```

text classification algorithms (Dai et al. 2007; Xue et al. 2008; Pan and Yang 2010). It contains nearly 20,000 newsgroup documents which have been evenly partitioned into 20 different newsgroups. As in the previous works (Dai et al. 2007; Xue et al. 2008), we generate six cross-domain text data sets from 20Newsgroups by utilizing its hierarchical structure. Specifically, the learning task is defined as the top-category binary classification, where our goal is to classify documents into one of the top-categories (e.g., *Comp*, *Rec*, etc.). For each data set, we select one top-category (e.g., *Comp*) as the positive class and another top-category (e.g., *Rec*) as the negative class. Then we select some subcategories (e.g., *comp.graphics* and *rec.motorcycles*) under the positive and the negative classes respectively to form a domain. In this work, we use the documents from four top-categories: *Comp*, *Rec*, *Sci* and *Talk* to generate data sets. Table 2 summarizes the data sets generated from 20Newsgroups.

Reuters-21578 The Reuters-21578 is another famous data set for evaluating text classification algorithms (Dai et al. 2007). As 20Newsgroups, the documents in Reuters-21578 are also organized with a hierarchical structure. For Reuters-21578, we use the preprocessed version of data sets provided in the web site (http://www.cse.ust.hk/TL/index.html) for experiments. This data set contains three cross-domain data sets which are generated with the documents from three biggest top-categories (i.e., *Orgs*, *People* and *Places*). Table 3 summarizes the generated data sets.

Table 2: Data Sets Generated from 20Newsgroups

Data set	Source Domain $\mathcal{D}^s$	Target Domain $\mathcal{D}^t$
Comp vs Rec	comp.graphics comp.sys.ibm.pc.hardware rec.motorcycles rec.sport.baseball	comp.os.ms-windows.misc comp.sys.mac.hardware rec.autos rec.sport.hockey
Comp vs Sci	comp.os.ms-windows.misc comp.sys.ibm.pc.hardware sci.electronics sci.space	comp.graphics comp.sys.mac.hardware sci.crypt sci.med
Comp vs Talk	comp.os.ms-windows.misc comp.sys.ibm.pc.hardware talk.politics.mideast talk.politics.misc	comp.graphics comp.sys.mac.hardware talk.politics.guns talk.religion.misc
Rec vs Sci	rec.autos rec.sport.baseball sci.crypt sci.med	rec.motorcycles rec.sport.hockey sci.electronics sci.space
Rec vs Talk	rec.autos rec.sport.baseball talk.politics.mideast talk.politics.misc	rec.motorcycles rec.sport.hockey talk.politics.guns talk.religion.misc
Sci vs Talk	sci.crypt sci.med talk.politics.misc talk.religion.misc	sci.electronics sci.space talk.politics.guns talk.politics.mideast

Table 3: Data Sets Generated from Reuters-21578

Data set	Source Domain $\mathcal{D}^s$	Target Domain $\mathcal{D}^t$
Orgs vs People	Orgs.{...}, People.{...}	Orgs.{...}, People.{...}
Orgs vs Places	Orgs.{...}, Places.{...}	orgs.{...}, Places.{...}
People vs Places	People.{...}, Places.{...}	People.{...}, Places.{...}

Baselines and Evaluation Criteria

To test the effectiveness of TCA, we compare it with two conventional classification algorithms: Support Vector Machine (SVM) and Logistic Regression (LG), and three state-of-the-art cross-domain classification methods: Spectral Feature Alignment (SFA) (Pan et al. 2010), Topic-bridge PLSA (TPLSA) (Xue et al. 2008) and Collaborative Dual-PLSA (CDPLSA) (Zhuang et al. 2010). For SVM and LG, the classifiers are trained with the labeled documents from the source domain and used to predict the class labels of unlabeled documents in the target domain. In SFA, the spectral clustering algorithm is adapted to co-cluster all words into the shared clusters for domain adaptation. Both TPLSA and CDPLSA aim to jointly model documents from different domains based on topic modeling. In TPLSA, all topics are assumed to be shared by different domains and used to represent documents. CDPLSA jointly models different domains by assuming that the associations between the topics and the document categories are stable across domains. In order to verify the usefulness of the induced feature mapping between domain-specific topics, we modify TCA by using only the shared topics to represent documents for classifier training. We denote it TCA$^{\text{share}}$. The classification accuracy is adopted as the evaluation criteria. For the algorithms which have the random initialization process, we conduct 10 and 50 random runs for the experiments on 20Newsgroups and Reuters-21578, respectively. And the average results of the random runs are reported.

Table 4: The Test Classification Accuracy on The Data Sets Generated from 20Newsgroups and Reuters-21578

Data Set	LG	SVM	SFA	TPLSA	CDPLSA	TCA$^{\text{share}}$	TCA
Comp vs Rec	0.906	0.895	**0.939**	0.910	0.914	0.867	**0.940**
Comp vs Sci	0.759	0.719	0.830	0.802	0.877	0.792	**0.891**
Comp vs Talk	0.911	0.898	**0.971**	0.938	0.955	0.912	0.967
Rec vs Sci	0.719	0.696	0.885	**0.928**	0.872	0.735	0.879
Rec vs Talk	0.848	0.827	0.935	0.849	0.912	0.828	**0.962**
Sci vs Talk	0.780	0.747	0.854	0.890	0.862	0.785	**0.940**
Orgs vs People	0.681	0.670	0.671	0.746	**0.808**	0.731	0.792
Orgs vs Places	0.692	0.669	0.683	0.719	0.714	0.660	**0.730**
People vs Places	0.513	0.520	0.506	0.623	0.548	0.614	**0.626**
Average	0.757	0.738	0.808	0.823	0.829	0.769	**0.859**

Table 5: Example of Domain-Specific Topics Extracted by Our Method on Data Set *Comp vs Sci*

Source Domain $\mathcal{D}^s$				Target Domain $\mathcal{D}^t$			
Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
windows	drive	edu	space	jpeg	mac	encryption	medical
dos	scsi	space	nasa	image	apple	government	disease
file	mb	henry	launch	file	db	clipper	health
com	ide	writes	earth	bit	edu	chip	cancer
edu	controller	nasa	spacecraft	gif	drive	law	patients
mouse	disk	toronto	orbit	color	scsi	key	hiv
os	bus	article	satellite	images	lc	security	treatment
ms	drives	pat	system	files	mb	privacy	vitamin
microsoft	hard	shuttle	solar	format	quadra	escrow	aids
win	dx	cost	data	quality	writes	nsa	infection

Implementation Details and Parameter Settings

For data preprocessing, we convert all words to lower cases and remove stop words. Besides, we filter out the words with document frequencies less than 3. For TCA, we set the total number of topics (i.e., $K + K^{\ell}$) in each domain to 12 and 20 for the experiments on 20Newsgroups and Reuters-21578, respectively. The topic numbers are tuned on some documents from data sets *Sci vs Talk* and *Orgs vs Places*. After tuning, the topic numbers are fixed and respectively used for the experiments on 20Newsgroups and Reuters-21578. The tuned documents are then put back to the original data sets. Without any prior knowledge, we simply set the proportion of the shared topics in each domain to 0.5, that is $K = K^{\ell}$. Since topic modeling methods can be sensitive to initial parameter values, we use the output of PLSA trained on individual domains to initialize the domain-specific topics, and the output of PLSA trained on the merged domain to initialize the shared topics. We set the hyperparameter α for Beta distribution to 20 in all experiments. The number of EM iterations is set to 200. LIBLINEAR LG (Fan et al. 2008) is used as the base classifier, and all parameters are set to their default values. For SFA, TPLSA and CDPLSA, we adopt the same parameter settings as the original papers.

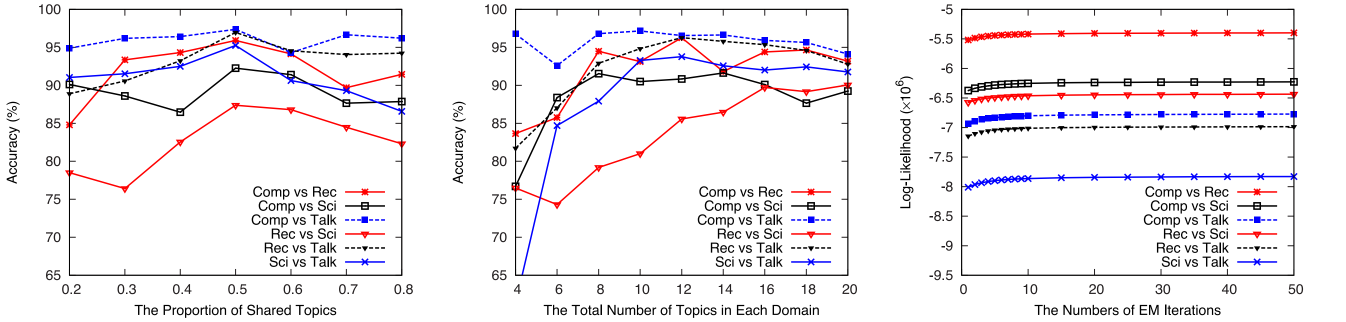
Experimental Results

In the first experiment, we compare our method with all baselines. Table 4 summarizes the classification performance on each data set. The last row of the table shows the average accuracy over all data sets. From the table, we can observe that our proposed method outperforms all baselines on six data sets. Table 5 presents the examples of extracted domain-specific topics on data set *Comp vs Sci*. We sort the words with the learned topic-word probability. The associated topic mapping matrix $\mathbf{U}$ is:

$$\mathbf{U} = \begin{bmatrix} 0.8046 & 0.4330 & -0.7232 & -0.3768 \\ 0.4348 & 0.9665 & -0.5876 & -0.4789 \\ -0.1364 & -0.3183 & 0.5664 & 0.1692 \\ -0.4292 & -0.4613 & 0.1722 & 0.1312 \end{bmatrix}$$

By examining the topical words and the topic mapping matrix, we can observe that the domain-specific topics having positive correlation scores are always semantically relevant. For example, Topic 1 in the source domain is about the Windows OS, and its positively correlated topics in the target domain (i.e., Topic 1 and Topic 2) are about the computer graphics and the Mac hardware, respectively. The result shows that our method can effectively identify the correlations between domain-specific features from different domains.

In the second experiment, we study the parameter sensitivity for the proportion of shared topics in each domain. In this experiment, we fix the total number of topics in each domain and vary the proportion of shared topics. Figure 2a shows the average classification accuracy of TCA under varying proportions of shared topics. We can observe that TCA performs well and steadily when the proportion of shared topics ranges from 0.4 to 0.6, which verifies the effectiveness of jointly modeling both the shared and the domain-specific topics.



(a) Classification accuracy under varying proportions of shared topics

(b) Classification accuracy under varying total numbers of topics in each domain

(c) Log-Likelihood under increasing numbers of EM Iterations

Figure 2: Parameter sensitivity analysis for: 1) The proportion of shared topics, 2) the total number of topics in each domain, and 3) the number of EM Iterations

In the third experiment, we study the parameter sensitivity for the total number of topics in each domain. In this experiment, we set the proportion of shared topics to 0.5 and vary the total numbers of topics in each domain. Figure 2b presents the performance of TCA under varying numbers of topics. As can be seen, TCA achieves good performance when the total number of topics in each domain is larger than 12.

In the last experiment, we test the convergence of the proposed Joint Mixture Model in TCA. Figure 2c shows the objective value under increasing number of EM iterations. We can observe that the log-likelihood grows quickly and converges after 50 iterations.

Conclusions

In this work, we propose a novel Topic Correlation Analysis (TCA) approach for cross-domain text classification. Unlike the previous works which focus on extracting only the shared latent features to bridge domains, in this paper, we show that the performance of cross-domain learning can be improved by further utilizing the domain-specific latent features, which remains unexplored. In the future, we intent to extend our work in the following directions: 1) Adapt the proposed method to the situation where a few documents are available in the target domain. 2) It will be interesting to study whether other effective methods, such as SCL, can be adopted to learn the correlations between topics.

Acknowledgment

The work was supported by National Natural Science Foundation of China (60973103,90924003) and HGJ National Key Project (2010ZX01042-002-002).

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *Proc. of EMNLP*, 120–128. Stroudsburg, PA, USA: ACL.

Breese, J. S.; Heckerman, D.; and Kadie, C. M. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. of UAI*, 43–52.

Dai, W.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2007. Co-clustering based classification for out-of-domain documents. In *Proc. of SIGKDD*, 210–219. New York, NY, USA: ACM.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *JMLR* 9:1871–1874.

Gupta, S. K.; Phung, D.; Adams, B.; Tran, T.; and Venkatesh, S. 2010. Nonnegative shared subspace learning and its application to social media retrieval. In *Proc. of KDD*, 1169–1178. New York, NY, USA: ACM.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proc. of SIGIR*, 50–57. New York, NY, USA: ACM.

Jiang, J., and Zhai, C. 2007. Instance weighting for domain adaptation in nlp. In *Proc. of ACL*, 264–271. Prague, Czech Republic: Association for Computational Linguistics.

Lin, J. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* 37(1):145 –151.

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

Pan, S. J.; Ni, X.; Sun, J.-T.; Yang, Q.; and Chen, Z. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proc. of WWW*, 751–760. New York, NY, USA: ACM.

Xue, G.-R.; Dai, W.; Yang, Q.; and Yu, Y. 2008. Topic-bridged plsa for cross-domain text classification. In *Proc. of SIGIR*, 627–634. New York, NY, USA: ACM.

Zhai, C.; Velivelli, A.; and Yu, B. 2004. A cross-collection mixture model for comparative text mining. In *Proc. of KDD*, 743–748. New York, NY, USA: ACM.

Zhuang, F.; Luo, P.; Shen, Z.; He, Q.; Xiong, Y.; Shi, Z.; and Xiong, H. 2010. Collaborative dual-plsa: mining distinction and commonality across multiple domains for text classification. In *Proc. of CIKM*, 359–368. New York, USA: ACM.