Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Domain Adaptation via Tree Kernel Based Maximum Mean Discrepancy for User Consumption Intention Identification

Xiao Ding, Bibo Cai, Ting Liu* and **Qiankun Shi**
Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
{xding, bbcai, tliu, qkshi}@ir.hit.edu.cn

Abstract

Identifying user consumption intention from social media is of great interests to downstream applications. Since such task is domain-dependent, deep neural networks have been applied to learn transferable features for adapting models from a source domain to a target domain. A basic idea to solve this problem is reducing the distribution difference between the source domain and the target domain such that the transfer error can be bounded. However, the feature transferability drops dramatically in higher layers of deep neural networks with increasing domain discrepancy. Hence, previous work has to use a few target domain annotated data to train domain-specific layers. In this paper, we propose a deep transfer learning framework for consumption intention identification, to reduce the data bias and enhance the transferability in domain-specific layers. In our framework, the representation of the domain-specific layer is mapped to a reproducing kernel Hilbert space, where the mean embeddings of different domain distributions can be explicitly matched. By using an optimal tree kernel method for measuring the mean embedding matching, the domain discrepancy can be effectively reduced. The framework can learn transferable features in a completely unsupervised manner with statistical guarantees. Experimental results on five different domain datasets show that our approach dramatically outperforms state-of-the-art baselines, and it is general enough to be applied to more scenarios. The source code and datasets can be found at `http://ir.hit.edu.cn/\%7exding/index_english.htm`.

1 Introduction

People used to express their needs and requirements on social media platforms in natural languages. One kind of such contents may explicitly or implicitly indicate the user wants to buy some specific products or services, which we refer to as user consumption intention. For example, "*Planning to buy an iPhone X, is it worth*?" implies that the user may want to buy "*iPhone X*". Ding et al. [2015] showed that consumption intentions are domain-dependent.

For such practical task, the performance of supervised learning models with insufficient training data cannot be satisfactory for applications, while manual annotating of sufficient training data for all domains is an almost impossible mission. Hence, a fruitful line of prior work explores transfer learning technology to reduce the human labeling efforts, by leveraging sufficient annotated data from the relevant source domain to the unlabeled target domain. The goal of domain adaptation is to transfer invariant structures or features from the source domain to the target domain, such that the distribution discrepancy of different domains can be alleviated. There are two main challenges for this task. On the one hand, how to reduce domain discrepancy places the first hurdle in adaptation models across domains. On the other hand, how to learn transferable feature representations would be a major bottleneck for the designing of domain adaptation mechanism.

A number of effective methods have been proposed to address the first challenge, by learning domain-invariant shallow features or re-weighting instances to reduce domain discrepancy. However, this line of work can only exploit explicit feature transferring and needs some labeled data to train the target domain model, which has been shown its limitation on domain adaptation [Long *et al.*, 2015]. Hence, Pan et al. [2008] proposed exploring a latent space, which can minimize the distance between distributions of the data in different domains, so that even when the target domain has no labeled data, they can still achieve a good performance. Pan et al. [2008] used Maximum Mean Discrepancy (MMD) [Huang *et al.*, 2007] to estimate the distance between different distributions, which has been shown more powerful than Kullback-Leibler (KL) divergence [Cao *et al.*, 2013] and Bregman divergence [Si *et al.*, 2010]. However, the limitation of MMD is that it depends on nonlinear kernel mapping to compute distribution discrepancy, and kernel-based MMD may suffer from vanishing gradients for low-bandwidth kernels. Some frequently-used kernels may not capture very complex distances in high dimensional spaces such as natural languages. Hence, kernel selection for specific tasks, especially for natural language processing (NLP) tasks is very important.

Recently, deep neural network has shown its power of learning transferable feature representations for cross do-

*Corresponding author

mains [Long *et al.*, 2017]. Deep neural network can learn representations that capture underlying factors, a subset of which may be relevant for each particular domain [Bengio *et al.*, 2013]. However, deep features are transited from general to specific along the network, and feature transferability drops dramatically in higher layers with increasing domain discrepancy. It means that the features represented by higher layers are domain-specific and cannot be directly transferred to other domains [Yosinski *et al.*, 2014]. Hence, previous work can only transfer features represented by lower layers in the network and has to use annotated data to train parameters of higher domain-specific layers [Oquab *et al.*, 2014].

In this paper, we propose an MMD-powered deep neural network framework for the task of domain adaptive consumption intention identification. In the framework, for lower layers, parameters can be directly transferred from the source domain to the target domain. For higher domain-specific layers, we use MMD to compare distributions of the source domain and the target domain in a reproducing kernel Hilbert space (RKHS). According to the MMD theory [Borgwardt *et al.*, 2006], the distance between distributions of two domain data is equivalent to the distance between the means of the two domain data mapped into an RKHS. The goal of our model is to train parameters of the higher layers such that the distributions of the source and target domain data can be close to each other. In other words, we want to find a feature map to minimize the distance between distributions of the source domain and the target domain. As the performance of two sample matching based on MMD is subject to the kernel choices, we employ a tree kernel method to reduce the domain discrepancy. Tree kernel function is widely used in NLP community [Culotta and Sorensen, 2004], as it can well capture the syntactic and semantic structures of sentences. In this paper, we model the tree structures of input sentences based on Tree-LSTM [Tai *et al.*, 2015].

The major contributions of the work presented in this paper are summarized as follows.

- We propose an MMD-powered deep neural network framework for the task of domain adaptive consumption intention identification, in which features of all layers can be transferred from the source domain to the target domain in a completely unsupervised manner.
- We explore tree kernel method for adapting deep representations, which substantially improves the performance of MMD in NLP tasks.
- Comprehensive experiments show that our approach can achieve better performance compared with state-of-the-art baseline methods evaluated on five domain adaptation datasets. We believe that the generalization ability of the proposed framework is strong enough to be applied to more NLP tasks.

2 Background

2.1 Consumption Intention Identification

Social media has become a popular venue for individuals to express their needs and desires. Such data can provide important resources for companies to identify users' consumption intentions, so that well-selected products or services can be recommended. A post with consumption intention means that it can explicitly or implicitly indicate that the individual may want to purchase some specific products or services [Ding *et al.*, 2015]. For example, a real tweet "*Finally decided to buy a new computer.*" explicitly indicates that the individual wants to buy a "*computer*". Meanwhile, the post "*Winter is coming, time to change clothes.*" does not explicitly state what product the individual wants to buy. But according to the semantic analysis of the sentence, we can speculate that the individual may want to buy "*winter clothes*".

Domain	Positive Instance	Negative Instance
M	Does anyone wanna go watch the fate of the furious with me today?	Be Fast and Furious if you wanna watch the fate of the furious.
A	I want to buy an air ticket to Stockholm.	Don't dare to fly!
T	Help me check if there is a train for Beijing this evening.	How do I get to Beijing south railway station?
P	Planning to buy an iPhone X, is it worth?	I don't want to use I9300.
C	Finally decided to buy a new computer.	How to install Windows 10 for my computer?

Table 1: The examples of positive and negative instances in five domain datasets.

Given an arbitrary domain sentence, our task is to identify whether it contains consumption intention or not. To this end, we build an MMD-based domain adaptation deep neural network framework to transfer knowledge from the source domain to the target domain in a completely unsupervised manner. There are five domain datasets: movie (**M**), booking air tickets (**A**), booking train tickets (**T**), phone (**P**), computer (**C**), for evaluating the effectiveness of our proposed model. Table 1 lists the examples of positive and negative instances in these domain datasets.

2.2 Maximum Mean Discrepancy in Reproducing Kernel Hilbert Space

The main challenge of domain adaptation is to estimate and minimize the difference in the probability distributions between the source domain and the target domain. There are a number of methods to estimate the distance between different distributions, such as KL divergence and Bregman divergence. However, these criteria are parametric as they require an intermediate density estimate. To approach this problem, we use MMD as a nonparametric estimation of the distance between different distributions, which compares distributions based on reproducing kernel Hilbert space. More specifically, we explore statistical tests of the null hypothesis that these distributions are equal, against the alternative hypothesis that these distributions are different. In this paper, we propose using the tree kernel variant of MMD (TK-MMD) to maximize two-sample test power and minimize the Type II error (the distributions are equal that is accepted despite the underlying distributions being different).

Formally, given two distributions p and q from the source domain and the target domain, respectively, we test whether

p and q are different on the basis of samples $\mathcal{X}^s = \{x_1^s, x_2^s, ..., x_m^s\}$ and $\mathcal{Z}^t = \{z_1^t, z_2^t, ..., z_n^t\}$ independently and identically distributed ($i.i.d.$) drawn from each of p and q, respectively, in a reproducing kernel Hilbert space $\mathcal{H}$. MMD is a kernel two-sample test, which accepts or rejects the null hypothesis $p = q$ based on the generated samples from different distributions. MMD is defined as follows:

$$MMD[p, q] \triangleq \sup_{\|f\|_{\mathcal{H}} \leq 1} (\mathbb{E}_{\mathcal{X}^s \sim p}[f(\mathcal{X}^s)] - \mathbb{E}_{\mathcal{Z}^t \sim q}[f(\mathcal{Z}^t)]) \quad (1)$$

$$MMD[\mathcal{X}^s, \mathcal{Z}^t] \triangleq \| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i^s) - \frac{1}{n} \sum_{j=1}^{n} \phi(z_j^t) \| \quad (2)$$

where $\mathcal{H}$ is a class of functions and $\phi(\cdot)$ is the nonlinear feature mapping that induces $\mathcal{H}$. The mean embedding of distribution p in $\mathcal{H}$ is a unique element $\mu_{\mathcal{X}}^s(p)$ such that $\mathbb{E}_{\mathcal{X}^s \sim p} f(\mathcal{X}^s) = \langle f(\mathcal{X}^s), \mu_{\mathcal{X}}^s(p) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. $\mathcal{H}$ must be rich enough to distinguish any two distributions and MMD can be defined as the distance between their mean embeddings: $MMD[p, q] = \|\mu_{\mathcal{X}^s}(p) - \mu_{\mathcal{Z}^t}(q)\|_{\mathcal{H}}^2$, so that $p = q$ if and only if $MMD[p, q] = 0$ [Gretton *et al.*, 2012].

The feature mapping takes the canonical form $\phi(x) = k(x^s, \cdot)$ [Steinwart and Christmann, 2008], where $k(x^s, z^t) : \mathcal{X}^s \times \mathcal{Z}^t \to \mathbb{R}$ is positive definite, and the notation $k(x^s, \cdot)$ indicates the kernel has one argument at x and the second is free. Given that in an RKHS, we can obtain the square MMD, $\|\mu_{\mathcal{X}^s}(p) - \mu_{\mathcal{Z}^t}(q)\|_{\mathcal{H}}^2$, in terms of kernel functions, and a corresponding unbiased finite sample estimate. In this paper, we use an estimate of the MMD to compare the square distance between the empirical kernel mean embeddings as

$$\begin{aligned} \widehat{MMD}[p, q] = & \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i^s, x_j^s) \\ & - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i^s, z_j^t) \\ & + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(z_i^t, z_j^t) \end{aligned} \quad (3)$$

where $\widehat{MMD}[p, q]$ is an unbiased estimator of $MMD[p, q]$.

The goal for this paper is to find an abstract feature representation through which the source and target domains are similar [Ben-David *et al.*, 2007]. To this end, we explore tree kernel based MMD to enhance the transferability of feature representation in neural network. Tree structure can well capture the semantic information of sentences. Hence, TK-MMD can better distinguish different distributions in RKHS for NLP tasks, so that it can strengthen the domain adaptation power and reduce the domain adaptation errors.

3 Domain Adaptive Consumption Intention Identification Model

In this paper, we formulate consumption intention identification as a classification problem and propose a **d**omain **a**daptive **c**onsumption **i**ntention identification model (DACI) in a completely unsupervised manner to solve it. In an unsupervised transfer learning setting, sufficient annotated data are available in the source domain $\mathcal{D}_{src} = \{(x_1^s, y_1^s), \ldots, (x_m^s, y_m^s)\}$, where $x_i^s \in \mathbb{R}^m$ is the input and y_i^s is the corresponding label to indicate whether the input contains consumption intention or not. While only unannotated data are available in the target domain $\mathcal{D}_{tar} = \{z_1^t, \ldots, z_n^t\}$, where $z_j^t \in \mathbb{R}^n$. Our task is to learn transferable features that can bridge cross-domain discrepancy and predict the labels y_j^t in the target domain using source domain supervision, based on a deep neural network framework.

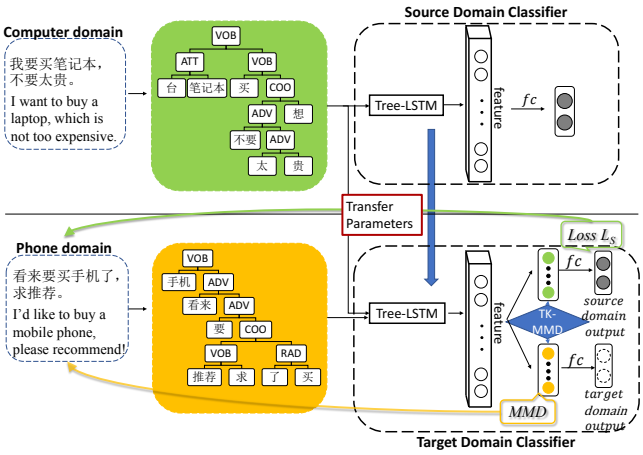

Figure 1: The architecture of DACI.

As shown in Figure 1, the framework consists of two steps. We first use Tree-LSTM [Tai *et al.*, 2015] to learn tree structured feature representations for the input sentences, and the parameters of feature representation layers can be directly transferred from the source domain to the target domain. Then we use a fully connected layer to connect the feature representation layer and the output layer. The parameters of the fully connected layer should be transferred via TK-MMD. In the following sections, we introduce each step in detail.

3.1 Feature Representation Learning

Exploring tree structures of sentences has been proven effective for NLP tasks. In this section, we start with introducing how to use Tree-LSTM [Tai *et al.*, 2015] to learn feature representations. We use Dependency Tree-LSTM, as dependency parsers can help us understand the relations between words and it is crucial for the task of consumption intention identification. The Tree-LSTM can be formulated by the following set of equations.

$$\begin{aligned} i_t &= \sigma(W^{(i)} x_t + U^{(i)} h_{t-1} + b^{(i)}), \\ f_t &= 1 - i_t, \\ o_t &= \sigma(W^{(o)} x_t + U^{(o)} h_{t-1} + b^{(o)}), \\ u_t &= tanh(W^{(u)} x_t + U^{(u)} h_{t-1} + b^{(u)}), \\ c_t &= i_t \odot u_t + f_t \odot c_{t-1}, \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \quad (4)$$

The c_t is the cell state and h_t is the hidden state. The $W^{(*)}$ and $U^{(*)}$ are the weight parameters, and $b^{(*)}$ is the bias. $\odot$ denotes the element-wise multiplication.

We regard the output vector in the last hop as the feature, and feed it to a *softmax* layer for classification. The model

is trained in a supervised manner by minimizing the cross entropy error of consumption intention classification, whose loss function is given below, where C means all training instances, L is the label set of whether the input sentence contains intention, (x^s, y^s) means a sentence-label pair.

$$J(\theta) = -\sum_{(x^s,y^s)\in C}\sum_{l\in L} p_l^g(x^s, y^s)\cdot log(p_l(x^s, y^s)) + \frac{\lambda}{2}\|\theta\|^2 \quad (5)$$

$p_l(x^s, y^s)$ is the probability of predicting (x^s, y^s) as category l produced by our model. $p_l^g(x^s, y^s)$ is 1 or 0, indicating whether the correct answer is l. θ stands for the parameters of Tree-LSTM. λ is an L2 regularization hyperparameter. We use back propagation to calculate the gradients of all the parameters, and update them with stochastic gradient descent.

3.2 Maximum Mean Discrepancy

Tree-LSTM is effective for the domain with sufficient training data. The main challenge is that the target domain has no annotated data, and hence directly adapting Tree-LSTM to the target domain is impossible. In standard LSTM, deep features eventually transit from general to specific by the last layer (fully connected layer) of the network, and the transferability gap increases with the domain discrepancy [Yosinski *et al.*, 2014]. Specifically, the fully connected (*fc*) layer is tailored to their source domain disregarding its performance on the target domain, hence it cannot be directly transferred to the target domain. In this paper, we first train Tree-LSTM on the source domain annotated data, and then make the distributions of the source and target to become similar under the mathing of the hidden representations of the *fc* layer. This can be realized by adding a TK-MMD-based adaptation regularizer function (2) to the Tree-LSTM loss function (5):

$$J'(\theta) = -\sum_{(x^s,y^s)\in C}\sum_{l\in L} p_l^g(x^s, y^s)\cdot log(p_l(x^s, y^s)) + \frac{\lambda}{2}\|\theta\|^2 + \gamma MMD[\mathcal{X}^s, \mathcal{Z}^t] \quad (6)$$

where $\gamma > 0$ is a penalty parameter. In our settings, parameters of feature representation layers in Tree-LSTM can be directly transferred from the source domain to the target domain. The parameters of the *fc* layer should be transferred via TK-MMD. The goal of our model is to find the parameters in Tree-LSTM that can minimize $J'(\theta)$.

3.3 Training

As shown in Eq. (3), the computation complexity of MMD^2 is $O(m+n)^2$, which may limit the efficiency of Tree-LSTM learning from large-scale data. Gretton et al. [2012] proposed a linear complexity estimation of MMD^2 instead. When $m = n$, a slightly simpler empirical estimate may be used. Let $U := (u_1, \ldots, u_m)$ be m *i.i.d* random variables, where $u := (x^s, z^t) \sim p \times q$ (i.e., x^s and z^t are independent). An unbiased estimate of MMD^2 is

$$MMD^2[p, q] = \frac{1}{m(m-1)}\sum_{i\neq j}^{m} h(u_i, u_j) \quad (7)$$

where $h(u_i, u_j) := k(x_i^s, x_j^s) + k(z_i^t, z_j^t) - k(x_i^s, z_j^t) - k(x_j^s, z_i^t)$.

Domain	Training (#instances)	Development (#instances)	Test (#instances)
M	4,432	554	558
A	10,064	1,258	1,268
T	7,807	976	990
P	7,088	886	892
C	6,016	752	754

Table 2: Statistics of datasets.

We train Tree-LSTM with mini-batch SGD, and compute the gradient of loss function Eq. (6) as $\bigtriangledown_\theta = \frac{\partial J(\mathbf{u})}{\partial \theta} + \gamma \frac{\partial h(\mathbf{u})}{\partial \theta}$. Such a mini-batch SGD can be easily implemented. We use Gaussian kernel in this paper.

4 Experiments

We compare DACI with state-of-the-art transfer learning baselines on five different domain datasets, focusing on the efficiency of adaption with tree-kernel MMD.

4.1 Data Description

We collected a large-scale microblog corpus for training initial word embedding from Sina Weibo. The corpus contains 20 million posts, 76 million sentences and 1.3 billion words.

To our knowledge, there is no public corpus for evaluating the task of domain adaptive consumption intention identification. Hence, we constructed a manually annotated sub corpus. As mentioned in Section 2, to evaluate the transferability of our model, we annotated five domain datasets (Phone domain datasets are from [Zhao *et al.*, 2014]). Detail statistics of training, development and test sets are shown in Table 2. For all sentences in our data, two annotators are asked to annotate whether it contains user consumption intention.

The agreement between our two annotators, measured using Cohen's Kappa Coefficient [Cohen, 1968], is substantial (kappa = 0.74 for consumption intention identification). We evaluate our model across the 20 transfer tasks: **M** $\rightarrow$ **A**, **M** $\rightarrow$ **T**, **M** $\rightarrow$ **P**, **M** $\rightarrow$ **C**, **A** $\rightarrow$ **M**, **A** $\rightarrow$ **T**, and so on.

4.2 Baseline Methods

We compare with the following baseline methods.

1. SCL [Blitzer *et al.*, 2007]: Conventional domain adaptation model uses structural corresponding learning on the task of sentiment classification.
2. SFA [Pan *et al.*, 2010]: Conventional domain adaptation model uses spectral feature alignment on the task of sentiment classification.
3. CNN [Ding *et al.*, 2015]: The low and middle layers' parameters of CNN are directly transferred from the source domain to the target domain, the high layer's parameters are fine-tuned by little annotated target domain data.
4. Tree-LSTM [Tai *et al.*, 2015]: Transfer parameters of all layers from the source domain to the target domain, to show the effectiveness of MMD.
5. CNN + GRL [Ganin and Lempitsky, 2015]: CNN is used as the basic consumption intention classification model,

Method	M	A	T	P	C
CNN	87.99	94.32	91.82	90.25	90.98
Tree-LSTM	89.43	94.16	97.87	93.61	94.83
DACI	76.52	91.48	88.38	91.03	91.64

Table 4: Best transferring results *vs.* Experimental results of CNN and Tree-LSTM model trained on the sufficient annotated data without transferring.

transferring deep features by augmenting CNN with a new gradient adaptation layer.

6. CNN+MMD [Long *et al.*, 2016]: CNN is used as the basic consumption intention classification model, and MMD is used to transfer task-specific layer's parameters from the source domain to the target domain.
7. LSTM+MMD: LSTM is used as the basic consumption intention classification model, to show the effectiveness of Tree-LSTM.

4.3 Results and Discussion

We first train DACI on the source domain, and then apply it to the target domain. Table 3 shows the classification accuracy (%) of the baseline systems as well as our methods on consumption intention identification. From the experimental results, we can make the following observations.

(1) Comparing with SFA and SCL, DACI achieved dramatically better performance, suggesting that deep neural network based methods can outperform conventional transfer learning methods by a large margin. This is mainly because deep neural networks can learn feature representations that capture underlying factors, a subset of which may be relevant for several particular domains [Bengio *et al.*, 2013].

(2) DACI also outperformed CNN and Tree-LSTM, which directly transfers features from the source domain to the target domain. This confirms that while general features can generalize to other domains, specific features tailored to the domain cannot help bridge the domain discrepancy. MMD can effectively minimize domain discrepancy via mean-embedding matching of the multi-layer representations across domains in a reproducing kernel Hilbert space.

(3) DACI achieved better performance than LSTM+MMD and CNN+MMD, which indicates that tree structured feature representations and tree kernel based MMD are more effective for enhancing feature transferability in NLP tasks, especially in consumption intention identification.

(4) Comparing between DACI and CNN+GRL, DACI also achieved better performance. The main reason is that MMD-based DACI can directly and explicitly reduce the domain discrepancy between the source domain and the target domain, compared to CNN+GRL.

(5) Most baseline methods cannot achieve consistently good performance over all transferring pairs. This is mainly because there is a large domain discrepancy between some domains. People may express their needs with different language patterns in different domains, which exacerbates domain discrepancy. Hence, baseline methods can achieve good performance on transferring pairs with small domain discrepancy, such as T $\rightarrow$ A and P $\rightarrow$ C; but the performance significantly drops with large domain discrepancy, such as M $\rightarrow$ A, M $\rightarrow$ T. Note that DACI can dramatically boost the performance on all transferring pairs.

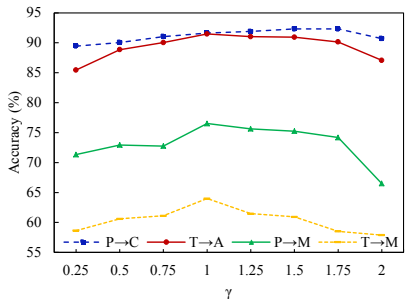

Figure 2: Accuracy vs. γ

To show the effectiveness of our method, we also compare with CNN and Tree-LSTM model trained on the sufficient annotated data without transferring. Experimental results (accuracy) are shown in Table 4. We only record the best transferring results of DACI in Table 4. The following observations can be made. (1) In some domains (e.g., A, P and C), unsupervised domain adaptation method can achieve comparative performance with supervised model. (2) Tree-LSTM outperforms CNN, which implies that Tree-LSTM can capture more semantic information of input sentences than CNN.

Parameter Sensitivity

We investigate the effects of the parameter γ. Figure 2 illustrates the variation of DACI's performance as $\gamma \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$ on tasks: P$\rightarrow$C, P$\rightarrow$M, T$\rightarrow$A, and T$\rightarrow$M. We find that the accuracy first increases and then decreases as γ varies. This confirms that a good trade-off between learning deep features and adapting distribution discrepancy can enhance feature transferability.

5 Related Work

5.1 Consumption Intention Identification

Identifying online users' commercial intents has attracted much attention [Dai *et al.*, 2006; Hollerit *et al.*, 2013; Zhao *et al.*, 2014; Wang *et al.*, 2015; Lo *et al.*, 2016]. Dai et al. [2006] first proposed to identify search queries that contain online commercial intention. Recently, social media has opened a new avenue for research on the consumption intention identification and a lot of interesting problems have been proposed. Yang and Li [2013] investigated the use of social media data to identify fundamental needs for individuals through a crowd-sourced study. Wang et al. [2013] mainly focused on identifying trend-driven commercial intents from microblogs. Hollerit et al. [2013] proposed to identify tweet-level commercial intents and link buyers and sellers. Ding et al. [2015] developed a consumption intention mining model based on convolutional neural network, for identifying whether the user has a consumption intention cross different domains. Lo et al. [2016] studied user activity and purchasing behavior with the goal of building models of time-varying user purchasing intent. Agarwal and Sureka [2017] conducted study on Tumblr microblogging website

Method	M→A	M→T	M→P	M→C	A→M	A→T	A→P
SCL [Blitzer *et al.*, 2007]	49.78	49.39	51.52	48.94	47.85	74.54	48.20
SFA [Pan *et al.*, 2010]	49.82	49.39	51.13	48.10	50.00	74.44	47.97
CNN [Ding *et al.*, 2015]	51.97	56.46	50.11	44.43	49.64	82.93	49.66
Tree-LSTM [Tai *et al.*, 2015]	47.87	48.08	56.50	49.87	49.46	86.97	49.44
CNN+GRL [Ganin and Lempitsky, 2015]	49.61	53.33	51.79	47.88	49.82	85.45	50.00
CNN+MMD [Long *et al.*, 2016]	55.13	68.89	49.79	48.01	52.58	84.85	53.42
LSTM+MMD	61.71	69.14	70.87	71.09	62.84	83.55	69.40
DACI(Ours)	**66.01**	**70.00**	**83.74**	**78.78**	**67.69**	**88.38**	**86.66**
Method	A→C	T→M	T→A	T→P	T→C	P→M	P→A
SCL [Blitzer *et al.*, 2007]	46.56	48.45	73.25	48.03	46.37	50.00	49.90
SFA [Pan *et al.*, 2010]	46.50	48.45	73.29	48.14	46.50	50.00	49.82
CNN [Ding *et al.*, 2015]	49.60	49.82	90.62	49.66	49.73	49.10	50.24
Tree-LSTM [Tai *et al.*, 2015]	49.34	50.00	90.38	49.89	50.00	49.82	49.13
CNN+GRL [Ganin and Lempitsky, 2015]	49.87	50.00	91.48	49.89	49.87	50.72	51.42
CNN+MMD [Long *et al.*, 2016]	49.47	52.05	89.51	51.91	50.00	51.39	49.13
LSTM+MMD	75.33	62.36	88.84	71.54	76.66	74.55	66.22
DACI(Ours)	**88.06**	**64.00**	**91.48**	**87.56**	**78.51**	**76.52**	**84.62**
Method	P→T	P→C	C→M	C→A	C→T	C→P	Avg
SCL [Blitzer *et al.*, 2007]	50.00	74.34	49.91	49.86	49.95	55.92	53.14
SFA [Pan *et al.*, 2010]	50.00	65.72	49.91	49.86	49.95	57.84	52.84
CNN [Ding *et al.*, 2015]	50.00	88.20	49.46	50.79	50.51	86.88	57.52
Tree-LSTM [Tai *et al.*, 2015]	49.90	91.25	50.36	50.47	49.90	86.66	57.76
CNN+GRL [Ganin and Lempitsky, 2015]	50.00	88.20	50.54	49.45	50.10	88.56	57.85
CNN+MMD [Long *et al.*, 2016]	52.42	88.86	50.00	57.09	50.51	86.43	59.66
LSTM+MMD	63.45	82.76	68.91	64.11	76.35	82.79	72.12
DACI(Ours)	**76.77**	**91.64**	**73.84**	**65.93**	**78.59**	**91.03**	**79.49**

Table 3: Experimental results on 20 domain transferring pairs, the best results are in bold.

and developed a cascaded ensemble learning classifier for identifying the posts having racist or radicalized intent. Our work adds to this line of work by performing a domain adaptive consumption intention identification model via tree kernel based maximum mean discrepancy. In contrast to previous work, we are among the first to address domain adaptation problem for consumption intention identification in a completely unsupervised manner.

5.2 Transfer Learning with Deep Neural Networks

The main challenge of transfer learning is how to reduce the discrepancy in data distributions across domains. Previous work learns a shallow representation model by which domain discrepancy is minimized, which cannot well capture underlying shared factors of variations [Pan *et al.*, 2008]. Deep neural networks can learn abstract representations that disentangle semantic information behind text data and extract transferable factors underlying different scenarios [LeCun *et al.*, 2015]. Nevertheless, deep neural networks can only reduce, but not remove, the cross-domain discrepancy [Yosinski *et al.*, 2014]. A fruitful line of prior work on deep domain adaptation explores effective domain discrepancy measurement and matching methods to boost transfer performance [Ding *et al.*, 2015]. However, deep features eventually transit from general to specific along the network, and feature transferability drops dramatically in higher layers with increasing domain discrepancy. Hence, Long et al. [2015] employed MMD to enhance the feature transferability of high layers in deep neural networks. This line of work mainly focuses on the area of computational vision. Our work adds to this line of work by using a more effective deep neural network – Tree-LSTM in NLP community to learn tree structured feature representations, and exploring TK-MMD to enhance the feature transferability of fully connected layers in LSTM.

6 Conclusion

In this paper, we proposed a domain adaptive neural network architecture to identify user consumption intention in social media across domains. The architecture takes the benefits of both deep neural network and two-sample test. We show that general features from lower layers of deep neural network can be well transferred, while specific features from higher layers may intensify domain discrepancy. Feature transferability can be enhanced by TK-MMD in a reproducing kernel Hilbert space. Experiments on five different domain datasets show that our model can achieve dramatically better performance compared with state-of-the-art baseline methods. Our proposed framework is general enough to be applied to other NLP tasks, such as sentiment analysis. In the future, we will apply our framework to more scenarios.

Acknowledgments

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Basic Research Program (973 Program) of China via Grant 2014CB340503, the National Natural Science Foundation of China (NSFC) via Grant 61472107 and 61702137. We are grateful to Wayne Xin Zhao for sharing annotated positive instances of Phone domain with us.

References

[Agarwal and Sureka, 2017] Swati Agarwal and Ashish Sureka. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint*, 2017.

[Ben-David *et al.*, 2007] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, 2007.

[Bengio *et al.*, 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2013.

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, Fernando Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[Borgwardt *et al.*, 2006] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 2006.

[Cao *et al.*, 2013] Xudong Cao, David Wipf, Fang Wen, Genquan Duan, and Jian Sun. A practical transfer learning algorithm for face verification. In *ICCV*, 2013.

[Cohen, 1968] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 1968.

[Culotta and Sorensen, 2004] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *ACL*, 2004.

[Dai *et al.*, 2006] Honghua Kathy Dai, Lingzhi Zhao, Zaiqing Nie, Ji-Rong Wen, Lee Wang, and Ying Li. Detecting online commercial intention (oci). In *WWW*, 2006.

[Ding *et al.*, 2015] Xiao Ding, Ting Liu, Junwen Duan, and Jian-Yun Nie. Mining user consumption intention from social media using domain adaptive convolutional neural network. In *AAAI*, 2015.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

[Hollerit *et al.*, 2013] Bernd Hollerit, Mark Kröll, and Markus Strohmaier. Towards linking buyers and sellers: detecting commercial intent on twitter. In *WWW*, 2013.

[Huang *et al.*, 2007] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[Lo *et al.*, 2016] Caroline Lo, Dan Frankowski, and Jure Leskovec. Understanding behaviors that lead to purchasing: A case study of pinterest. In *KDD*, 2016.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[Long *et al.*, 2016] Mingsheng Long, Jianmin Wang, Yue Cao, Jiaguang Sun, and S Yu Philip. Deep learning of transferable representation for scalable domain adaptation. *TKDE*, 2016.

[Long *et al.*, 2017] Mingsheng Long, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. 2017.

[Oquab *et al.*, 2014] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[Pan *et al.*, 2008] Sinno Jialin Pan, James T Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *AAAI*, 2008.

[Pan *et al.*, 2010] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW*, 2010.

[Si *et al.*, 2010] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *TKDE*, 2010.

[Steinwart and Christmann, 2008] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[Tai *et al.*, 2015] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, 2015.

[Wang *et al.*, 2013] Jinpeng Wang, Wayne Xin Zhao, Haitian Wei, Hongfei Yan, and Xiaoming Li. Mining new business opportunities: Identifying trend related products by leveraging commercial intents from microblogs. In *EMNLP*, 2013.

[Wang *et al.*, 2015] Jinpeng Wang, Gao Cong, Wayne Xin Zhao, and Xiaoming Li. Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. In *AAAI*, 2015.

[Yang and Li, 2013] Huahai Yang and Yunyao Li. Identifying user needs from social media. *IBM Research Division, San Jose*, 2013.

[Yosinski *et al.*, 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014.

[Zhao *et al.*, 2014] Xin Wayne Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *KDD*, 2014.