Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis

[12]Chunning Du, [12]Haifeng Sun, [12]Jingyu Wang,* [12]Qi Qi, [12]Jianxin Liao
[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]EBUPT Information Technology Co., Ltd., Beijing 100191, China
{duchunning,sunhaifeng_1,wangjingyu,qiqi}@ebupt.com

Abstract

Cross-domain sentiment classification aims to address the lack of massive amounts of labeled data. It demands to predict sentiment polarity on a target domain utilizing a classifier learned from a source domain. In this paper, we investigate how to efficiently apply the pre-training language model BERT on the unsupervised domain adaptation. Due to the pre-training task and corpus, BERT is task-agnostic, which lacks domain awareness and can not distinguish the characteristic of source and target domain when transferring knowledge. To tackle these problems, we design a post-training procedure, which contains the target domain masked language model task and a novel domain-distinguish pre-training task. The post-training procedure will encourage BERT to be domain-aware and distill the domain-specific features in a self-supervised way. Based on this, we could then conduct the adversarial training to derive the enhanced domain-invariant features. Extensive experiments on Amazon dataset show that our model outperforms state-of-the-art methods by a large margin. The ablation study demonstrates that the remarkable improvement is not only from BERT but also from our method.

1 Introduction

Sentiment analysis aims to automatically identify the sentiment polarity of the textual data. It is an essential task in natural language processing with widespread applications, such as movie reviews and product recommendations. Recently, deep networks have significantly improved the state-of-the-art in sentiment analysis. However, training deep networks is highly depended on a large amount of labeled training data which is time-consuming and requires expensive manual labeling. Thus, there is a strong need to leverage or reuse rich labeled data from a different but related source domain. Cross-domain sentiment analysis, which transfers the knowledge learned from source domain to a new target domain, becomes a promising direction.

The main challenge of cross-domain sentiment analysis is domain discrepancy due to different expression of the user's emotion across domains. To address the problem, a wide-used approach is designed to extract domain invariant features, which means that the distributions of features from the source domain and target domain are similar (Zellinger et al., 2017; Persello and Bruzzone, 2016; Ganin et al., 2016; Yu and Jiang, 2016a). One effective way to obtain domain-invariant features is adversarial training(Ganin et al., 2016; Li et al., 2017; Zheng et al., 2019). Specifically, A domain discriminator is learned by minimizing the classification error of distinguishing the source from the target domains, while a deep classification model learns transferable representations that are indistinguishable by the domain discriminator.

Very recently, pre-trained language models have shown to be effective for improving many language tasks (Peters et al., 2018). Bidirectional Encoder Representations from Transformers (BERT) realized a breakthrough, which pre-trains its encoder using language modeling and by discriminating surrounding sentences in a document from random ones (Devlin et al., 2019). Pre-training in this manner can construct bidirectional contextual representation, and the large-scale pre-training enables BERT powerful ability in language understanding. We only need to add one output layer and fine-tune BERT to get the state-of-the-art results in sentiment analysis. Theoretically, BERT can enhance the performance in cross-domain sentiment analysis. However, some important problems remain when directly fine-tuning BERT in the task of cross-domain sentiment analysis:

Firstly, there is no labeled data in the target do-

*Corresponding author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4019–4028
July 5 - 10, 2020.

main which brings many difficulties to the fine-tune procedure. If we fine-tune BERT only by the source labeled data, the shift between training and test distributions will degrade the BERT's performance. Secondly, BERT is task-agnostic and has almost no understanding of opinion text. BERT is pre-trained by the universal language Wikipedia, leaving the domain challenges unresolved (Xu et al., 2019). For example, in the pre-training procedure, BERT may learn to guess the [MASK] in "The [MASK] is bright" as "sun". But in a laptop sentiment analysis, it is more likely to be "screen". Especially, in the cross-domain sentiment analysis scenario, the labeled data is limited, which is insufficient to fine-tune BERT to ensure full domain-awareness. Thirdly, cross-domain sentiment analysis also arises a new challenge for BERT to distinguish the characteristic of source and target domain to keep the transferable features and abandon domain-specific information.

To address above problems, we design a novel pre-training task to make BERT domain-aware and then improve the BERT's fine-tuning procedure by adversarial training. Specifically, a novel post-training procedure is implemented that adapts BERT with unlabeled data from different domains to enhance the domain-awareness. Apart from the target domain masked language model task, we introduce the domain-distinguish pre-training task. BERT will be pre-trained to distinguish whether the two sentences come from the same domain. The domain-distinguish pre-training task will encourage BERT to distill syntactic and semantic domain-specific features, so as to be domain-aware. The proposed post-training procedure gives us a new way to fully utilize language knowledge from the target domain and link the source and target in a self-supervised way. Based on this, we could then conduct the adversarial training to derive the enhanced domain-invariant features.

Experiments on Amazon reviews benchmark dataset show that our model gets the average result 90.12% in accuracy, 4.22% absolute improvement compared with state-of-the-art methods. The ablation study shows that the remarkable improvement is not only from BERT but also from our method. The contributions of this paper can be summarized as follows:

- We apply BERT to cross-domain sentiment analysis task and leverage the post-training method to inject the target domain knowledge to BERT.
- A novel domain-distinguish pre-training task is proposed for the BERT post-training. This design encourages BERT to be domain-aware and distill the domain-specific features in a self-supervised way.

2 Related Work

2.1 Cross-Domain Sentiment Analysis

Cross-domain sentiment analysis aims to generalize a classifier that is trained on a source domain, for which typically plenty of labeled data is available, to a target domain, for which labeled data is scarce. There are many pivot-based methods (Blitzer et al., 2007a; Yu and Jiang, 2016b; Ziser and Reichart, 2018; Peng et al., 2018), which focus on inducing a low-dimensional feature representation shared across domains based on the co-occurrence between pivots and non-pivots. However, selecting pivot words is very tedious, and the pivot words are manually selected, which may not be accurate. Recently, some adversarial learning methods (Ganin et al., 2016; Li et al., 2017; Zheng et al., 2019) propose to train the feature generator to minimize the classification loss and simultaneously deceive the discriminator, which is end-to-end without manually selecting pivots.

2.2 Language Model Pre-training

Pre-trained language representations with self-supervised objectives have become standard in a wide range of NLP tasks. Previous work can be divided into two main categories: feature-based approaches and fine-tuning approaches.

The recent proposed feature-based approaches mainly focus on learning contextualized word representations such as CoVe (McCann et al., 2017) and ELMo (Peters et al., 2018). As with traditional word embeddings, these learned representations are also typically used as features in a downstream model. On the other hand, fine-tuning approaches mainly pre-train a language model on a large corpus with an unsupervised objective and then fine-tune the model with in-domain labeled data for downstream applications. The advantage of these approaches is that few parameters need to be learned from scratch. Specifically, Howard and Ruder (2018) propose ULMFiT, which uses a different learning rate for each layer with learning warmup and gradual unfreezing. GPT (Radford et al., 2018) uses a transformer encoder (Vaswani

et al., 2017) instead of an LSTM and jointly fine-tunes with the language modeling objective. Moreover, BERT (Devlin et al., 2019) is a large-scale language model consisting of multiple layers of transformer, which further incorporates bidirectional representations. BERT is the state-of-art pre-training language model. However, in the cross-domain sentiment analysis scenario, BERT is task-agnostic and can not distinguish the characteristic of source and target domain.

3 Model

In this section, we introduce the proposed model for cross-domain sentiment analysis in detail. We begin by giving the problem definition and notations. Then BERT and post-training method are formally presented in the second subsection. Finally, the adversarial training process is introduced. We also give a theoretical analysis of our model.

3.1 Problem Definition and Notations

In the task of cross-domain sentiment analysis, we are given two domains D_s and D_t which denote a source domain and a target domain, respectively. In the source domain, $D_s^l = \{x_s^i, y_s^i\}_{i=1}^{N_s^l}$ are N_s^l labeled source domain examples, where x_s^i means a sentence and y_s^i is the corresponding polarity label. There are also N_s^u unlabeled data $D_s^u = \{x_s^i\}_{i=1+N_s^l}^{N_s^l+N_s^u}$ in the source domain. In the target domain, there is a set of unlabeled data $D_t = \{x_t^i\}_{i=1}^{N_t}$, where N_t is the number of unlabeled data. Cross-domain sentiment analysis demands us to learn a robust classifier trained on labeled source domain data to predict the polarity of unlabeled sentences from the target domain.

3.2 Background of BERT

BERT (Devlin et al., 2019) builds on the Transformer networks (Vaswani et al., 2017), which relies purely on attention mechanism and allows modeling of dependencies without regard to their distance in the input sequences. BERT is pre-trained by predicting randomly masked words in the input (MLM task) and classifying whether the sentences are continuous or not (NSP task). The MLM task allows the word representation to fuse the left and the right context, and the NSP task enables BERT to infer the sentences' relationship. The pre-trained BERT can be easily fine-tuned by one softmax output layer for classification task.

3.3 BERT Post-training

Despite the success, BERT suffers from the domain challenge. BERT is pre-trained by Wikipedia, leading to task-agnostic and little understanding of opinion text. Especially in the cross-domain sentiment analysis scenario, the lack of abundant labeled data limits the fine-tune procedure, which degrades BERT due to the domain shift. This task also demands BERT to distinguish the characteristic of source and target domain for better knowledge transfer. Therefore, we propose BERT post-training which takes BERT's pre-trained weights as the initialization for basic language understanding and adapts BERT by novel self-supervised pre-trained tasks: domain-distinguish task and target domain masked language model.

3.3.1 Domain-distinguish Task

The next sentence prediction (NSP) task encourages BERT to model the relationship between sentences beyond word-level, which benefits the task of Question Answering and Natural Language Inference. However, cross-domain sentiment analysis operates on a single text sentence and does not require the inference ability. Instead, the ability to distinguish domains plays an important role. Therefore, during the post-training procedure, we replace the NSP task by domain-distinguish task (DDT). Specifically, we construct the sentence-pair input: `[CLS] sentence A [SEP] sentence B [SEP]`, where `[CLS]` and `[SEP]` are special embeddings for classification and separating sentences. 50% of time `sentence A` and `sentence B` are all randomly sampled from target domain reviews, we label it `TargetDomain`. And 50% of time `sentence A` and `sentence B` come from target domain and another domain, whose label is `MixDomain`. We do not fix the collocation, in another word, we only ensure that the two sentences come from different domains but the order is random. For example:

Input = `[CLS] The mouse is smooth and great [SEP] The screen is plain [SEP]`

Label = `TargetDomain`

Input = `[CLS] This book is boring [SEP] The system of the laptop is stable [SEP]`

Label = `MixDomain`

The domain-distinguish pre-training is a classifi-

cation task. We add one output layer on the pooled representation and maximize the likelihood of the right label. The domain-distinguish pre-training enables BERT to distill the specific features for different domains, which enhances the downstream adversarial training and benefits the cross-domain sentiment analysis.

3.3.2 Target Domain MLM

To inject the target domain knowledge, we leverage the masked language model (MLM) (Devlin et al., 2019). It requires to predict the randomly masked words in the sentence, which encourages BERT to construct a deep bidirectional representation. In the cross-domain sentiment analysis, there are no labeled data but plenty of unlabeled data in the target domain to post-train BERT by MLM. Specifically, we replace 15% of tokens by `[MASK]` at random. The final hidden vectors corresponding to the mask tokens are fed into an output softmax over the vocabulary. We maximize the likelihood of the masked token id.

Post-training by unlabeled review data in target domain will effectively alleviate the shift of domain knowledge. For example, if the masked word is an opinion word in "This movie is `[MASK]`", this objective challenges BERT to learn the representation for fine-grained opinion words in movie review domain, such as "touchable" or "disturbing".

One problem is that the DDT task mixes sentences from other domains in the sentence pair. The sentence from other domains will be the noise which brings domain bias. Therefore, we only mask the tokens in target domain sentences if the domain-distinguish task label is `MixDomain`.

The total loss of the post-training procedure is the sum of losses of target domain MLM and domain-distinguish task. The adaptation takes about 5 hours to complete on one single NVIDIA P100 GPU.

3.4 Adversarial Training

The post-training procedure injects target domain knowledge and brings domain-awareness to BERT. Based on the post-trained BERT, we now could utilize the adversarial training to abandon the distilled domain-specific features to derive the domain-invariant features. Specifically, a sentiment classifier and a domain discriminator is designed operating on the hidden state $h_{[CLS]}$ of the special classification embedding `[CLS]`.

3.4.1 Sentiment Classifier

The sentiment classifier is simply a fully-connected layer and outputs the probabilities through a softmax layer:

$$y_s = \text{softmax}(W_s h_{[CLS]} + b_s). \quad (1)$$

The classifier is trained by the labeled data in the source domain and the loss function is cross-entropy:

$$L_{sen} = -\frac{1}{N_s^l} \sum_{i=1}^{N_s^l} \sum_{j=1}^{K} \hat{y}_s^i(j) \log y_s^i(j), \quad (2)$$

where $\hat{y}_s^i \in \{0, 1\}$ is the ground truth label in the source domain, and K denotes the number of different polarities.

3.4.2 Domain Discriminator

The domain discriminator aims to predict domain labels of samples, i.e., coming from the source or target domain. The parameters of BERT are optimized to maximize the loss of the domain discriminator. This target will encourage BERT to fool the domain discriminator to generate domain-invariant features.

Specifically, before feeding $h_{[CLS]}$ to the domain discriminator, the hidden state of classification embedding `[CLS]` $h_{[CLS]}$ goes through the gradient reversal layer (GRL) (Ganin et al., 2016). During the forward propagation, the GRL acts as an identity function but during the backpropagation, the GRL reverses the gradient by multiplying it by a negative scalar λ. GRL can be formulated as a 'pseudo-function' $Q_\lambda(x)$ by two equations below in order to describe its forward- and backward-behaviors:

$$Q_\lambda(x) = x, \quad (3)$$

$$\frac{\partial Q_\lambda(x)}{\partial x} = -\lambda I. \quad (4)$$

We denote the hidden state $h_{[CLS]}$ through the GRL as $Q_\lambda(h_{[CLS]}) = \hat{h}_{[CLS]}$ and then feed it to the domain discriminator as:

$$d = \text{softmax}(W_d \hat{h}_{[CLS]} + b_d). \quad (5)$$

The target is to minimize the cross-entropy for all data from the source and target domains:

$$L_{dom} = -\frac{1}{N_s + N_t} \sum_{i}^{N_s+N_t} \sum_{j}^{K} \hat{d}^i(j) \log d^i(j), \quad (6)$$

where $\hat{d}^i \in \{0, 1\}$ is the ground truth domain label. Due to the GRL, the parameters for domain discriminator θ_{dd} are optimized to increase the ability to predict domain labels, however, the parameters for BERT θ_{BERT} are optimized to fool the domain discriminator, leading to domain-invariant features.

3.4.3 Joint Learning

The sentiment classifier and the domain discriminator are jointly trained, and the total loss is:

$$L_{total} = L_{sen} + L_{dom}. \tag{7}$$

The post-training procedure and our proposed domain-distinguish pre-training task will enhance the adversarial training to obtain lower classification error in the target domain, we will analyze it in Sec 3.5.

3.5 Theoretical Analysis

In this section, we provide a theoretical analysis of our approach. First, we provide an insight into existing theory, then we introduce an expansion of the theory related to our method and explain how the post-training and adversarial training cooperate to obtain a remarkably better result than state-of-the-art methods.

For each domain, there is a labeling function on inputs X, defined as $f : X \to [0, 1]$. The ideal label functions for source and target domain are denoted as: f_s and f_t, respectively. We define a hypothesis label function h: $X \to [0, 1]$ and a disagreement function:

$$\epsilon(h_1, h_2) = E[|h_1(x) - h_2(x)|]. \tag{8}$$

Then the expected error on the source samples of h is defined as: $\epsilon_s(h) = \epsilon_s(h, f_s)$. For the target domain, we have: $\epsilon_t(h) = \epsilon_t(h, f_t)$.

The divergence between source and target domain could thus be measured by $\mathcal{H}\Delta\mathcal{H}$-distance, which is defined as follows:

$$d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D_t) = 2 \sup_{h,h' \in \mathcal{H}} |\epsilon_s(h, h') - \epsilon_t(h, h')| \tag{9}$$

This distance is firstly proposed in (Ben-David et al., 2010) and frequently used to measure the adaptability between different domains (Shen et al., 2018; Chen et al., 2019).

3.5.1 Theorem 1.

Let H be the hypothesis class. Given two different domains D_s, D_t, we have:

$$\forall h \in H, \epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D_t) + C \tag{10}$$

This theorem means that the expected error on the target domain is upper bounded by three terms: (1) the expected error on the source domain; (2) the divergence between the distributions D_s and D_t; (3) the error of the ideal joint hypothesis. Normally, C is disregarded because it is considered to be negligibly small. Therefore, the first and second terms are important quantitatively in computing the target error.

For the first term, the error rate of source domain ϵ_s, it is easy to minimize with source labeled training data. Moreover, we adopt BERT, which brings powerful contextual representation for lower error rate. The second item in Eq. 10 demands us to generate similar features among different domains. Our proposed domain-distinguish pre-training task and post-training for BERT enable the model to identify the specific features for different domains. This ability will enhance the domain discriminator, which will help to find more complicated domain specific features and get abandoned by adversarial training. Therefore, we further decrease the divergence between the domains, which is quantitatively measured by $\mathcal{A}$-distance in Sec 4.6.

4 Experiments

In this section, we empirically evaluate the performance of our proposed methods.

4.1 Datasets and Experimental Setting

We conduct the experiments on the widely-used Amazon reviews benchmark datasets collected by (Blitzer et al., 2007b). It contains reviews from four different domains: Books (B), DVDs (D), Electronics (E) and Kitchen appliances (K). For each domain, there are 2,000 labeled reviews and approximately 4000 unlabeled reviews. Following the convention of previous works (Ziser and Reichart, 2018; Ganin et al., 2016; Qu et al., 2019), we construct 12 cross-domain sentiment analysis tasks. For each task, we employ a 5-fold cross-validation protocol, that is, in each fold, 1600 balanced samples are randomly selected from the labeled data for training and the rest 400 for validation.

S → T	Previous Models					BERT				
	DANN	PBLM	HATN	ACAN	IATN	BERT	HATN-BERT	BERT-AT	BERT-DA	BERT-DAAT
D → B	81.70	82.50	86.30	82.35	87.00	89.40	89.81	89.55	90.40	**90.86**
E → B	78.55	71.40	81.00	79.75	81.80	86.50	87.10	87.15	88.31	**88.91**
K → B	79.25	74.20	83.30	80.80	84.70	87.55	87.88	87.65	87.90	**87.98**
B → D	82.30	84.20	86.10	83.45	86.80	88.96	89.36	89.70	**89.75**	89.70
E → D	79.70	75.00	84.00	81.75	84.10	87.95	88.81	88.20	89.03	**90.13**
K → D	80.45	79.80	84.50	82.10	84.10	87.30	87.89	87.72	88.35	**88.81**
B → E	77.60	77.60	85.70	81.20	86.50	86.15	87.21	87.30	88.11	**89.57**
D → E	79.70	79.60	85.60	82.80	86.90	86.55	86.99	86.05	88.15	**89.30**
K → E	86.65	87.10	87.00	86.60	87.60	90.45	90.31	90.25	90.59	**91.72**
B → K	76.10	82.50	85.20	83.05	85.90	89.05	89.41	89.55	90.65	**90.75**
D → K	77.35	83.20	86.20	78.60	85.80	87.53	87.59	87.69	88.55	**90.50**
E → K	83.95	87.80	87.90	83.35	88.70	91.60	92.01	91.91	92.75	**93.18**
Average	80.29	80.40	85.10	82.15	85.90	88.25	88.69	88.56	89.37	**90.12**

Table 1: Accuracy of domain adaptation on Amazon benchmark.

4.2 Implementation Details

We adopt $\text{BERT}_{\text{base}}$(uncased) as the basis for all experiments. When generating the post-training data, each sentence in the target domain gets duplicated 10 times with different masks and sentences pair. We limit the maximum sequence length is 256. During the post-training, we train with batch size of 16 for 10000 steps. The optimizer is Adam with learning rate 2e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01. During the adversarial training, The weights in sentiment classifier and domain discriminator are initialized from a truncated normal distribution with mean 0.0 and stddev 0.02. In the gradient reversal layer (GRL), we define the training progress as $p = \frac{t}{T}$, where t and T are current training step and the maximum training step, respectively, and the adaptation rate λ is increased as $\lambda = \frac{2}{1+\exp(-10p)} - 1$.

4.3 Compared Methods

We compare our method with 5 state-of-the-art methods: **DANN** (Ganin et al., 2016), **PBLM** (Ziser and Reichart, 2018), **HATN** (Li et al., 2018), **ACAN** (Qu et al., 2019), **IATN** (Zhang et al., 2019). We also design several variants of BERT as baselines:

- **BERT**: Fine-tuning vanilla BERT by the source domain labeled data.
- **HATN-BERT**: HATN (Li et al., 2018) model based on BERT.
- **BERT-AT**: This method conducts the adversarial training operating on vanilla BERT.
- **BERT-DA**: Fine-tuning domain-aware BERT by the source domain labeled data. The domain-aware BERT is obtained by post-training.
- **BERT-DAAT**: Our proposed method introduced in Sec 3.

4.4 Experimental Results

Table 2 shows the classification accuracy of different methods. We can observe that the proposed BERT-DAAT outperforms all other methods.

For the previous models, they mostly base on the word2vec (Mikolov et al., 2013) or glove (Pennington et al., 2014). Compared to BERT's contextual word representation, they can not model complex characteristics of word use and how these uses vary across linguistic contexts, resulting in relatively worse overall performance. We can see that the vanilla BERT, which is fine-tuned only by the source domain labeled data without utilizing target domain data, can still outperform all the previous methods. For fair comparison, we reproduce the experiment of HATN model (Li et al., 2018) that incorporates BERT as the base model. As shown in Table 2, HATN-BERT achieves a comparable result with BERT-AT.

For the BERT variants, we did not see a remarkable improvement in the results of BERT-AT, which conducts adversarial training on BERT. It demonstrates that, in the task of cross-domain sentiment analysis, the bottleneck of BERT is the lack of domain-awareness and can not be tackled purely

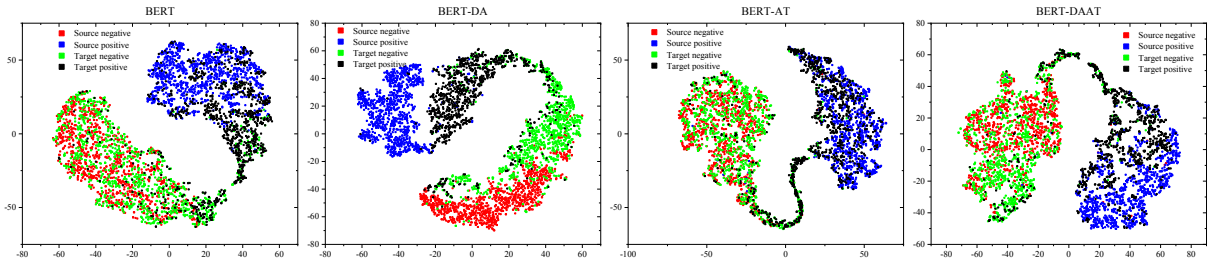

Figure 1: The effect of post-training and adversarial training on the distribution of the extracted features. The figure shows t-SNE visualization of the BERT's hidden state for the B → E task. The red, blue, green and black points denote the source negative, source positive, target negative and target positive examples, respectively.

by adversarial training. On the contrary, the post-training procedure could improve the result by 1.12% on average. It verifies the effectiveness of our proposed post-training methods that could inject the domain knowledge to BERT. As expected, BERT-DAAT performs best among the variants of BERT, 0.75% absolute improvement to BERT-DA and 1.87% absolute improvement to BERT, showing that the post-training procedure could further enhance the adversarial training.

4.5 Visualization of Features

To intuitively assess the effects of the post-training and adversarial training on BERT, we further perform a visualization of the feature representations of the variants of BERT for the training data in the source domain and the testing data in the target domain for the B → E task. As shown in Figure 1, the graphs are obtained by applying t-SNE on the set of all representation of source and target data points. Every sample is mapped into a 768-dimensional feature space through BERT and projected back into a two-dimensional plane by the t-SNE.

In the vanilla BERT representation (first subgraph in Figure 1), we could observe that data points of different polarities in source domain are well separated. While for the target domain, some data points are mixed together. It shows that only utilizing source domain labeled data is not enough for the target domain classification. For the post-trained BERT (subgraph for BERT-DA), data points belong to four clusters, indicating that domains and sentiment polarities are both well classified. It verifies that our post-training strategy brings domain-awareness to BERT. Moreover, compared to the first subgraph, the boundary for sentiment polarity classification is more clear, showing that injecting domain knowledge by post-training is beneficial to sentiment classification.

The latter two subgraphs in Figure 1 are the feature distributions obtained by adversarial training. One common characteristic is that data samples from different domains are very close to each other through adversarial training. However, the boundary for sentiment polarity classification is not very clear in BERT-AT's feature representation, resulting in degraded performance. For our proposed BERT-DAAT, the post-training enables the domain-awareness and help to distill more complicated domain specific features. The adversarial training is thus enhanced to get more domain-invariant features. We can find that target points are homogeneously spread out among source points, which decreases the divergence between the domains. According to Theorem 10, it can lower the upper boundary of the target error.

4.6 $\mathcal{A}$-distance

Theorem 10 shows that the divergence between domains $d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D_t)$ plays an important role. To quantitatively measure it, we compare the $\mathcal{A}$-distance, which is usually used to measure domain discrepancy (Ben-David et al., 2010). The definition of $\mathcal{A}$-distance is: $d_{\mathcal{A}} = 2(1 - 2\epsilon)$, where ϵ is the generalization error of a classifier trained with the binary classification task of discriminating the source domain and target domain. More precisely, to obtain $\mathcal{A}$-distance, we firstly split source and target domain data into two subsets of equal size and get the feature representation. We then train a linear SVM on the first subset to predict which domain the sample comes from. The error rate ϵ could be calculated on the second subset through the trained SVM, and $\mathcal{A}$-distance is obtained by $d_{\mathcal{A}} = 2(1 - 2\epsilon)$.

We compare the $\mathcal{A}$-distance of BERT, BERT-AT, and BERT-DAAT. Results are shown in Figure 2. For each cross-domain sentiment analysis task,

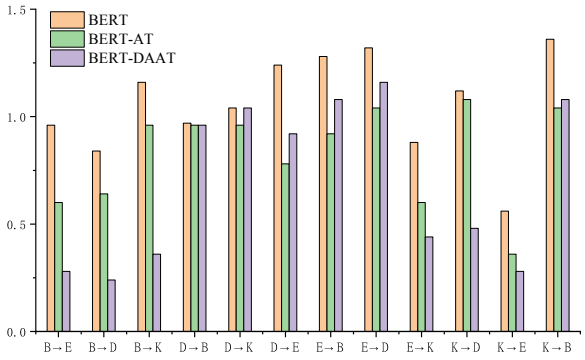

Figure 2: Comparison of $\mathcal{A}$-distance of different models.

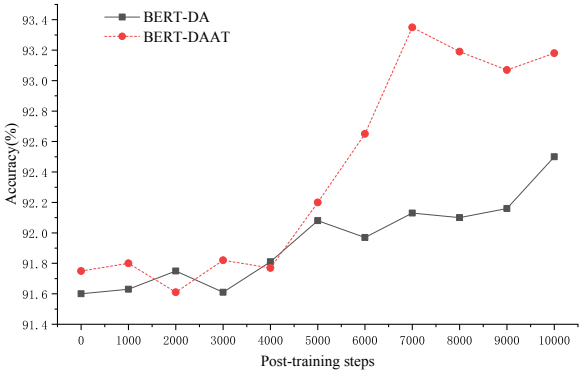

Figure 3: Ablation study on the number of post-training steps. The x-axis is the value of post-training steps k. The y-axis is the accuracy on the task of E → K.

the $\mathcal{A}$-distance of BERT is highest. It is easy to conclude that applying adversarial training can effectively decrease the $\mathcal{A}$-distance. Overall, the $\mathcal{A}$-distance of BERT-DAAT is lower than BERT-AT, verifying that the post-training could enhance the adversarial training to decrease the domain discrepancy.

4.7 Ablation Studies

To analyze the effect of different components including post-training steps and post-training tasks, we conduct the ablation experiments.

4.7.1 Effects of Post-Training Steps

In this subsection, we study the effect of post-training steps. Figure 3 presents the accuracy on the task of E→K based on the checkpoint that has been post-trained for k steps. The results for BERT-DA are obtained by fine-tuning source domain labeled data, BERT-DAAT is adversarial training by source labeled data and target unlabeled data.

We find that, with limited post-training steps (fewer than 5000 steps), BERT-DA and BERT-DAAT perform similarly with BERT and BERT-AT, respectively. However, given post-training steps more than 5000, both the results of BERT-DA and BERT-DAAT see an increase. Especially, after post-training more than 5000 steps, BERT-DAAT shows remarkable strengths compared to BERT-DA. This shows that plenty of post-training steps is necessary to inject domain knowledge and domain-awareness.

4.7.2 Effects of Post-training Tasks

The post-training tasks in our work include target domain masked language model (MLM) and our proposed domain-distinguish task (DDT). We design two models which ablate MLM and DDT separately and compare them with BERT-DAAT on the tasks of D→B, E→B, and K→B. Results in Table 2 indicate that: the target domain masked language model task (MLM) and domain-distinguish task(DDT) are both beneficial to cross-domain sentiment analysis.

Model	D→B	E→B	K→B
BERT	89.40	86.50	87.55
BERT-DAAT	**90.86**	**88.91**	**87.98**
-w/o MLM	89.91	87.39	87.80
-w/o DDT	90.02	88.01	87.63

Table 2: Ablation study over post-training tasks. w/o means without.

5 Conclusion and Future Work

In this paper, we propose the BERT-DAAT model for cross-domain sentiment analysis. Our purpose is to inject the target domain knowledge to BERT and encourage BERT to be domain-aware. Specifically, we conduct post-training and adversarial training. A novel domain-distinguish pre-training task is designed to distill the domain-specific features in a self-supervised. Experimental results on Amazon dataset demonstrate the effectiveness of our model, which remarkably outperforms state-of-the-art methods.

The proposed post-training procedure could also be applied to other domain adaptation scenarios such as named entity recognition, question answering, and reading comprehension. In the future, we would like to investigate the application of our theory in these domain adaptation tasks.

Acknowledgements

This work was supported in part by the National Key R&D Program of China 2018YFB1800502, in part by the National Natural Science Foundation of China under Grants 61671079 and 61771068, in part by the Beijing Municipal Natural Science Foundation under Grant 4182041, and in part by the Ministry of Education and China Mobile Joint Fund MCM20180101. This work was also supported by BUPT Excellent Ph.D. Students Foundation CX2020206.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*.

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

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

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. 2019. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL, 2018*.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *AAAI, 2018*.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*.

Minlong Peng, Qi Zhang, Yu-Gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP 2014*.

Claudio Persello and Lorenzo Bruzzone. 2016. Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning. *IEEE Trans. Geoscience and Remote Sensing*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*.

Xiaoye Qu, Zhikang Zou, Yu Cheng, Yang Yang, and Pan Zhou. 2019. Adversarial category alignment network for cross-domain sentiment classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. In *Technical report, OpenAI*.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *AAAI, 2018*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017,*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*.

Jianfei Yu and Jing Jiang. 2016a. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 236–246.

Jianfei Yu and Jing Jiang. 2016b. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *EMNLP, 2016*.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. In *ICLR, 2017*.

Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. 2019. Interactive attention transfer network for cross-domain sentiment classification. In *AAAI 2019*.

Li Zheng, Li Xin, Wei Ying, Bing Lidong, Zhang Yu, and Yang Qiang. 2019. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*.