An Adaptive Hybrid Framework for Cross-domain Aspect-based Sentiment Analysis

Yan Zhou[1], **Fuqing Zhu**[1*], **Pu Song**[1,2], **Jizhong Han**[1], **Tao Guo**[1], **Songlin Hu**[1,2*]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{zhouyan, zhufuqing, songpu, hanjizhong, guotao, husonglin}@iie.ac.cn

Abstract

Cross-domain aspect-based sentiment analysis aims to utilize the useful knowledge in a source domain to extract aspect terms and predict their sentiment polarities in a target domain. Recently, methods based on adversarial training have been applied to this task and achieved promising results. In such methods, both the source and target data are utilized to learn domain-invariant features through deceiving a domain discriminator. However, the task classifier is only trained on the source data, which causes the aspect and sentiment information lying in the target data can not be exploited by the task classifier. In this paper, we propose an Adaptive Hybrid Framework (AHF) for cross-domain aspect-based sentiment analysis. We integrate pseudo-label based semi-supervised learning and adversarial training in a unified network. Thus the target data can be used not only to align the features via the training of domain discriminator, but also to refine the task classifier. Furthermore, we design an adaptive mean teacher as the semi-supervised part of our network, which can mitigate the effects of noisy pseudo labels generated on the target data. We conduct experiments on four public datasets and the experimental results show that our framework significantly outperforms the state-of-the-art methods.

Introduction

Aspect-based sentiment analysis (ABSA) is an important task in fine-grained sentiment analysis (Liu 2012; Pontiki et al. 2014; Thelwall et al. 2010), which involves aspect extraction(AE) and aspect sentiment classification(ASC). Given a sentence, the goal of ABSA is to identify the aspect terms and infer the sentiment expressed on these aspects. For example, given a sentence "I love Windows 7 which is a vast improvement over Vista.", this task need to extract the aspect terms "Windows 7" and "Vista", and predict the sentiment polarities towards them are positive and negative respectively. Many supervised models have been proposed for this task (Li et al. 2019a; He et al. 2019; Zhou et al. 2019) and achieved promising results. Unfortunately, these models are highly dependent on large-scale training dataset which may be not available in many practical situations. One of the solutions to this problem is cross-domain aspect-based sentiment analysis, which utilizes the labeled data in a source domain to improve the prediction of the aspect terms with their sentiment polarities in a target domain.

There are some studies focused on the domain adaptation of fine-grained sentiment analysis. Ding et al. (Ding, Yu, and Jiang 2017) combined neural network with rule-based auxiliary task for the domain adaptation of aspect extraction, which is only a subtask of ABSA. Wang and Pan (Wang and Pan 2018) presented a recursive neural network for cross-domain aspect and opinion co-extraction, which utilized the shared dependency structure between different domains. To alleviate the dependency on linguistic resources, Wang and Pan (Wang and Pan 2019) used local and global memory units to capture the transferable interactions of aspect and opinion terms. These researches merely focused on the detection of aspect and opinion terms, while ABSA also need to find out the correspondences between them. Li et al. (Li et al. 2019b) proposed a transferable neural network with selective adversarial training for cross-domain ABSA, which is the first attempt for the domain adaptation of ABSA. In their model, the unlabeled target data are used by the domain discriminator to learn the common features across domains, while the task classifier is only trained on the source data. This leads the target data can not be utilized to refine the task classifier, even though these data may contain some beneficial aspect and sentiment information.

In this paper, we propose an Adaptive Hybrid Framework (AHF) which combines semi-supervised learning and adversarial training together for cross-domain aspect-based sentiment analysis. On the one hand, our framework utilizes the pseudo labels generated on the target data to train the task classifier, which can improve the task decision boundary. On the other hand, our framework is able to align the feature distribution through adversarial training. Specifically, AHF is comprised by a domain discriminator, a student network and a teacher network. The domain discriminator equipped with a gradient reversal layer is used to learn domain-invariant features. We employ Mean Teacher model (Tarvainen and Valpola 2017) which is comprised by a student and a teacher network to implement pseudo-label based semi-supervised learning. The teacher network tracks an exponential moving average of the student network weights, thus the teacher tends to generate better pseudo labels than the student. The student is trained on the labeled source data and the target data with pseudo labels.

*Corresponding author.

$\boldsymbol{x}$ (sentence)	I	love	Windows	7	which	is	a	vast	improvement	over	Vista	.
$\boldsymbol{y}$ (labels)	O	O	B-POS	I-POS	O	O	O	O	O	O	B-NEG	O

Table 1: An example sentence and its corresponding labels.

Though better pseudo labels can be produced by the teacher, there still exists a lot of noisy labels due to the domain discrepancy. To reduce the impact of the noisy labels, we further modify the original Mean Teacher with automatic threshold and domain similarity weighted loss, which is refered as adaptive mean teacher. First, different from the previous work using a same threshold (French, Mackiewicz, and Fisher 2018), we employ automatic confidence threshold to filter the noisy labels. It is unreasonable to employ a same threshold for our task, because the number of aspect terms is far less than non-aspect terms in the dataset and the network is prone to predict higher probability for non-aspect terms. Therefore we separately compute the thresholds for aspect terms and non-aspect terms at different training steps. In addition, we assign a higher weight to the label of a word, if the word comes from a sentence more similar with the source data.

The main contributions of this paper are as follows:

- For the first time, we propose to integrate pseudo-label based semi-supervised learning and adversarial learning for cross-domain ABSA. Therefore our framework AHF can effectively leverage the unlabeled target data for task classifier refinement and domain-invariant feature learning.
- To reduce the effects of noisy pseudo labels, we introduce an adaptive mean teacher network to implement the semi-supervised learning, which extends the original Mean Teacher with automatic confidence threshold and similarity weighted loss.
- Experimental results on ten transfer pairs show that our framework can achieve better performance than the state-of-the-art methods.

Related Work

Aspect-based Sentiment Analysis. Many models have been proposed to tackle the task of aspect-based sentiment analysis. Mitchell et al. (Mitchell et al. 2013) used hand-designed features and conditional random fields (CRFs) to detect the aspects and their sentiment polarities. With the development of deep learning, numerous neural network models have been applied to the task of ABSA. Zhang et al. (Zhang, Zhang, and Vo 2015) combined word embedding and automatic extracted features for this task. A unified framework comprised by three key components was proposed by Li et al. (Li et al. 2019a). He et al. (He et al. 2019) presented a novel interactive multi-task learning network which can leverage the interactions between the two subtasks of ABSA. Zhou et al. (Zhou et al. 2019) presented a span-based model to jointly extract the aspects and classify the sentiment expressed on them. These supervised models have shown promising results on this task, but they generally rely on numerous annotated data, which are hard to be collected in real world applications.

Cross-domain Aspect-based Sentiment Analysis. There have been a lot of researches for the domain adaptation of coarse-grained sentiment classification (Li et al. 2017; He et al. 2018; Li et al. 2018; Zhang et al. 2019). Howerver, coarse-grained sentiment classification only predicts an overall sentiment polarity of a document, while ABAS need to predict the aspects with their sentiments, which is more challenging. The tasks of cross-domain aspect or opinion terms extraction are closely related to cross-domain ABSA. Traditional approaches employed feature engineering or bootstrapping algorithm for the domain adaptation of aspect extraction (Jakob and Gurevych 2010; Li et al. 2012). Recent researches focused on neural network based models. Ding et al. (Ding, Yu, and Jiang 2017) integrated the neural network with rule-based unsupervised methods for cross-domain aspect extraction. Wang and Pan (Wang and Pan 2018) proposed a novel recursive neural network to the domain adaptation of aspect and opinion co-extraction. But these researches only perform aspect or opinion extraction without identifying their correspondences. Li et al. (Li et al. 2019b) published the first work for cross-domain aspect-based sentiment analysis. They proposed a selective adversarial learning method to align the features between different domains. Since the target data are merely used to train the domain discriminator, the task classifier can not leverage the category information lying in the target data.

Mean Teacher. The Mean Teacher model was proposed by Tarvainen et al. (Tarvainen and Valpola 2017) for semi-supervised image classification. Xu et al. (Xu et al. 2019) adopted this model for the domain adaptation of semantic segmentation. Recently, it was applied for the NLP tasks such as document classification (Ko, Durrett, and Li 2019) and semantic parsing (Wang et al. 2020). In this paper, we apply the Mean Teacher to a sequence labeling problem and modify it to fit our task.

Our Approach

In this section, we introduce the overall framework of our proposed approach. We first give the task definition and notations. And then we describe the architecture of AHF in detail.

Task Definition

We formulate the task of ABSA as a sequence labeling problem. Given an input sentence $\boldsymbol{x} = \{w_1, w_2, ..., w_n\}$, our goal is to predict a label sequence $\boldsymbol{y} = \{l_1, l_2, ..., l_n\}$, where each l_t is from the label set $\mathcal{Y}$ = {O, B-POS, I-POS, B-NEG, I-NEG, B-NEU, I-NEU}. Label "O" represents the word is not in an aspect term. Other labels are comprised by two parts: {B, I} denote the beginning and inside of an aspect term; {POS, NEG, NEU} denote the positive, negative and

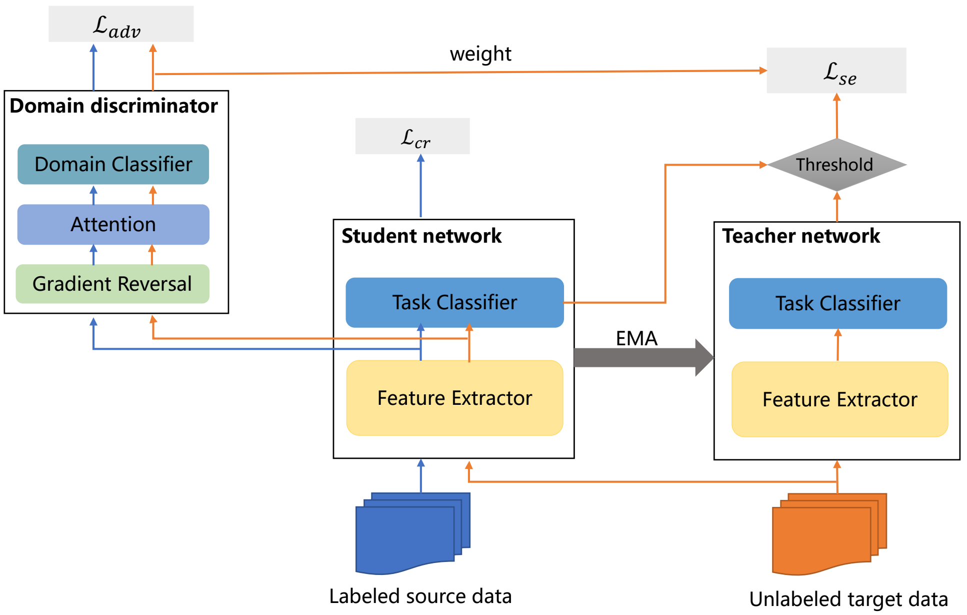


Figure 1: The architecture of AHF.

neutral sentiment. We present an example sentence and its corresponding label sequence in Table 1. From the labels of the words, we can get "Windows 7" and "Vista" are the aspect terms, and their sentiment polarities are positive and negative, respectively.

In this paper, we focus on the unsupervised domain adaptation of ABSA. The input data includes a set of labeled sentences from a source domain $\mathcal{D}_{\mathcal{S}} = \{\boldsymbol{x}_{\boldsymbol{s}}^{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{s}}^{\boldsymbol{i}}\}_{i=1}^{N_s}$ and a set of unlabeled sentences from a target domain $\mathcal{D}_{\mathcal{T}} = \left\{\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{j}}\right\}_{j=1}^{N_t}$. Our goal is to use both $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$ to train a model, which can predict the label sequences for target data.

An Overview of AHF

We present the architecture of AHF in Figure 1. It consists of a domain discriminator, a student network and a teacher network. The domain discriminator receives the vectors output from feature extractor of the student network. Then it predicts an input sentence is from the source or the target domain. And it is trained on the two domain data. The student network and the teacher network share a same structure and they constitute the adaptive mean teacher. The student network is trained with a cross entropy loss and a squared error loss. The source data with ground truth annotations are used to compute the standard cross entropy loss. For the unlabeled target data, we use the teacher network to generate the soft pseudo labels and these data are used to calculate the squared error loss. In particular, we employ automatic threshold to filter some noisy pseudo labels. And according to the sentence similarity with the source domain, we assign different weights to the labels of the words when computing squared error. The teacher is an ensembling network of the student, which does not participate in the back-propagation. At each time step, its parameters are updated through the exponential moving average (EMA) of the corresponding parameters in the student network.

Base Model

The student network and the teacher network share a same architecture and we introduce this architecture in this section. The base model is comprised by a feature extractor and a task classifier.

Feature Extractor. For an input sentence $\boldsymbol{x} = \{w_1, w_2, ..., w_n\}$, the model first converts each word into an embedding vector. We create the embedding vector e_t of word w_t through concatenating its word embedding ew_t and POS embedding ep_t:

$$e_t = [ew_t; ep_t] \tag{1}$$

where ";" represents the vector concatenation. Then we employ a BiLSTM to capture the contextual information of each word in the input sentence and obtain the contextual representation h_t as follows:

$$h_t = [\overrightarrow{LSTM}(e_t); \overleftarrow{LSTM}(e_t)]. \tag{2}$$

Task Classifier. The task classifier is comprised by a linear layer and a softmax layer. It uses the hidden vectors from BiLSTM $\boldsymbol{h} = \{h_1, h_2, ..., h_n\}$ as input and predicts the label for each word. The probability distribution y_t of the t-th word is computed as follows:

$$y_t = \text{softmax}(W_y h_t + b_y) \tag{3}$$

where W_y and b_y denote the weight matrix and bias term, respectively.

Domain Discriminator

To empower the feature extractors with the ability of learning domain-invariant features, we add a domain discriminator to implement the adversarial learning (Goodfellow et al.

2014). In particular, the domain discriminator utilizes all the sentences in $\mathcal{D_S}$ and $\mathcal{D_T}$ as inputs and attempts to tell which domain these sentences come from. At the same time, the feature extractor of the student network tries to fool the domain discriminator.

In detail, we use $\boldsymbol{hs} = \{hs_1, hs_2, ..., hs_n\}$ to represent the hidden vectors computed by the feature extractor of the student network. These hidden vectors are fed into the domain discriminator and encoded by an attention layer into a single vector. The attention layer first computes a weight α_t for each word to represent the importance of them:

$$m_t = V_m^T \tanh(W_m hs_t + b_m) \tag{4}$$

$$\alpha_t = \frac{\exp(m_t)}{\sum_{k=1}^{n} \exp(m_k)} \tag{5}$$

where V_m^T and W_m are weight matrices, b_m is the bias term. The representation of the sentence is computed by the weighted sum of all the hidden vectors:

$$r_i = \sum_{t=1}^{n} \alpha_t hs_t. \tag{6}$$

Finally, we adopt a softmax layer to do the domain classification:

$$d_i = \text{softmax}(W_d r_i + b_d). \tag{7}$$

We employ an adversarial loss $\mathcal{L}_{adv}$ to enforce the student feature extractor to produce domain shared representations. For each sentence $\boldsymbol{x}_i$, we set $z_i = 1$ if it is from the source domain, otherwise we set $z_i = 0$. The adversarial loss is computed as follows:

$$\mathcal{L}_{adv} = -\sum_{i=1}^{N_s+N_t} (z_i \log(d_i) + (1 - z_i)\log(1 - d_i)) \tag{8}$$

where N_s and N_t represent the number of the sentences in the source dataset $\mathcal{D_S}$ and the target dataset $\mathcal{D_T}$, respectively. The common training process is to minimize the classification error. But our intention is to learn the features such that the discriminator cannot recognize which domain a sentence comes from. To resolve this issue, we add a gradient reversal layer (Ganin et al. 2016) before the attention layer to conduct adversarial training. We use θ_f^S to denote the parameters in the student feature extractor. And the gradient reversal layer reverses $\frac{\partial \mathcal{L}_{adv}}{\partial \theta_f^S}$ into $-\eta \frac{\partial \mathcal{L}_{adv}}{\partial \theta_f^S}$ during the gradient backpropagation process. Thus the common features can be obtained through the training of domain discriminator.

Adaptive Mean Teacher

So far, the data distribution can be aligned through the domain discriminator. But the aspect and sentiment information lying in the target data are not explicitly utilized. A natural solution is to produce the pseudo labels for the target data and use these labels to guide the training of the task classifier. The effectiveness of this solution is highly dependent on the quality of the generated labels. And the predictions computed by an exponential moving average of the network tend to be more reliable than the ones produced by the original network (Tarvainen and Valpola 2017). Thus we employ the Mean Teacher to leverage the pseudo labels in our framework. Furthermore, we modify the Mean Teacher with automatic threshold and similarity weighted loss to reduce the influence of the noisy labels, which is referred as adaptive mean teacher.

The Mean Teacher consists of a student network and a teacher network, which have the same structure following the base model introduced above. The teacher does not participate in the back-propagation. Its parameters are updated with the exponential moving average of the corresponding parameters in the student network. At the time step p, the parameters of the teacher θ_p^T are computed as follows:

$$\theta_p^T = \gamma \theta_{(p-1)}^T + (1 - \gamma)\theta_p^S \tag{9}$$

where θ_p^S represents all the parameters in the student network at time step p, γ is a smoothing coefficient hyperparameter.

The source domain sentences are fed into the student network to predict the label of each word. We employ the cross entropy as the task loss function:

$$\mathcal{L}_{cr} = -\sum_{i=1}^{N_s} \sum_{t=1}^{N_i} g_{i,t} log(y_{i,t}^S) \tag{10}$$

where N_s denotes the number of the sentences in source dataset $\mathcal{D_S}$, N_i is the total word number in the i-th sentence, $g_{i,t}$ is the one-hot vector representing the gold label of the t-th word in the i-th sentence, $y_{i,t}^S$ is the probability distribution computed by the task classifier following the Equation 3. This loss function enables the student network to learn features from the labeled source data.

The target domain sentences are fed into both the student network and the teacher network. The teacher network generates soft pseudo labels for these sentences. At the same time, the student is trained to produce consistent predictions with these soft pseudo labels. In our framework, we employ automatic threshold and similarity weighted loss to mitigate the impact of noisy pseudo labels.

While high network prediction confidence does not guarantee correctness, there exists a positive correlation between them, so we first use the confidence threshold as filter. Previous work (French, Mackiewicz, and Fisher 2018) employed a fixed threshold throughout the training process, but this is not suitable for our task. The number of non-aspect terms is much more than the aspect terms, which leads that the network tends to predict higher probability value for label "O". Therefore we compute the thresholds for label "O" and aspect labels, respectively. We divide the label set $\mathcal{Y}$ into two sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$, where $\mathcal{Y}_1$ only contains label "O" and $\mathcal{Y}_2$ is comprised by other aspect labels. For each label set $\mathcal{Y}_c$ ($c \in \{1, 2\}$), we select $\beta\%$ most-reliable predictions by looking at network predictions on target data in the current batch. Then we use $\mathcal{Q}_\beta(\mathcal{Y}_c)$ to denote the smallest confidence value in these reliable predictions and compute the threshold τ_c of $\mathcal{Y}_c$ at the current training step as:

$$\tau_c = \max(\rho_c, \mathcal{Q}_\beta(\mathcal{Y}_c)) \tag{11}$$

where ρ_c is a manually set value for $\mathcal{Y}_c$, which is utilized to filter the noisy labels at the early training steps. In this way, the threshold is adaptive both to different label sets and to different training phases. For the t-th word in the i-th sentence, the probability distribution computed by the teacher task classifier is represented by $y_{i,t}^T$. The threshold τ_c is adopted to calculated the mask matrix $M_{i,t}$ as:

$$M_{i,t} = \begin{cases} 1, \text{ if } (\max(y_{i,t}^T) > \tau_c) \ \wedge \ (l_{i,t}^T \in \mathcal{Y}_c) \\ 0, \text{ otherwise} \end{cases} \qquad (12)$$

where $l_{i,t}^T$ represents the predicted label of the teacher network for the t-th word. Furthermore, we consider that the predicted labels are more reliable for the sentences more similar with the source domain. And the output probability d_i from the domain discriminator can be interpreted as a measure of the similarity. Thus, to effectively leverage the pseudo labels on the target data, we employ the squared error with threshold mask matrix and similarity weight as loss function:

$$\mathcal{L}_{se} = \sum_{i=1}^{N_t} \sum_{t=1}^{N_i} d_i \cdot M_{i,t} \cdot \| y_{i,t}^T - y_{i,t}^S \|^2 \qquad (13)$$

where N_t denotes the number of sentences in the target dataset $\mathcal{D}_\mathcal{T}$, $y_{i,t}^T$ and $y_{i,t}^S$ are the probability distribution computed by the teacher and the student network, respectively.

Joint Training

We combine the three losses into an overall objective function:

$$\mathcal{L} = \mathcal{L}_{cr} + \mathcal{L}_{se} + \lambda_{adv}\mathcal{L}_{adv} \qquad (14)$$

where λ_{adv} is a hyper-parameter. During the training phase, the network learns to minimize $\mathcal{L}$ with respect to the model parameters except the adversarial learning part which will be maximized. In the test phase, the sentences are fed into the teacher network to predict the labels.

Experiments

Datasets

In order to evaluate the performance of our model, we conduct experiments on four benchmark datasets: Restaurant(R), Laptop(L), Device(D) and Service(S). The datasets are from four different domains and the statistics of them are described in Table 2. The restaurant data is comprised by the restaurant reviews from SemEval 2014 (Pontiki et al. 2014), SemEval 2015 (Pontiki et al. 2015) and SemEval 2016 (Pontiki et al. 2016). The laptop data consists of laptop reviews in SemEval 2014 (Pontiki et al. 2014). The device data is created by Hu and Liu (Hu and Liu 2004) and contains the review sentences from five different digital products. The service datasets contains reviews from the web service and is introduced by Toprak et al. (Toprak, Jakob, and Gurevych 2010).

Dataset	Domain	#Sentences	#Train	#Test
R	Restaurant	3900	2481	1419
L	Laptop	1869	1458	411
D	Device	1437	954	483
S	Service	2153	1433	720

Table 2: The statistics of the datasets.

Following the recent work of cross-domian ABSA (Li et al. 2019b), we construct ten transfer pairs based on the four datasets mentioned above. Since the laptop domain and device domain are quite similar, we do not use the transfer pairs L → D and D → L. This paper focuses on the unsupervised domain adaptation of aspect-based sentiment analysis, and so there are only unlabeled data from the target domain. The training dataset of each transfer pair contains the labeled training data of the source domain and the unlabeled training data of the target domain. We use the labeled testing data of the source domain as the validation set. And the testing data of the target domain is used as the evaluation set. We employ the F1 score as evaluation metric in our experiments. The extracted results is considered to be correct only if both the words in the aspect terms and the sentiments towards the aspects are the same as the gold annotated span.

Experimental Settings

We use Stanford Parser (Manning et al. 2014) to generate part-of-speech tags of the sentences and employ word2vec tool[1] on two different corpora to get the word embeddings. One corpus contains 1M laptop domain reviews from Amazon (McAuley et al. 2015) and the other corpus is from Yelp Challenge dataset[2]. The dimensions of the word embedding and POS embedding are set to 100 and 15, respectively. We employ two layers BiLSTM in our experiment and the hidden units of LSTM is set to 100. We apply dropout over the embeddings layers and BiLSTM layers with the dropout rate 0.5. The parameters about adversarial learning η and λ_{adv} are set 1 and 0.1, respectively. The value of the smoothing coefficient parameter γ is 0.98. The values of β, ρ_1 and ρ_2 are set as 60, 0.9 and 0.4, respectively. We set the batch size of the source domain data and target domain data as 32. The parameters are optimized by RMSprop algorithm with learning rate 0.001. We run the experiments five times and report the average results.

In order to comprehensively evaluate our method, we compare our framework with several fine-grained adaptation methods:

- **TCRF**(Jakob and Gurevych 2010): A traditional sequence model combines CRF and hand engineered features such as POS tags, short dependency path and word distance.
- **RAP**(Li et al. 2012): A relational adaptive bootstrapping algorithm that uses common opinion words and syntactic relations to expand aspect and opinion lexicons.
- **Hier-Joint**(Ding, Yu, and Jiang 2017): Use LSTM and auxiliary labels generated by syntactic rules to extract cross-domain aspect terms.

[1] https://radimrehurek.com/gensim/models/word2vec.html

[2] http://www.yelp.com/dataset challenge

Model	D → R	D → S	L → R	L → S	R → D	R → L	R → S	S → D	S → L	S → R	Avg
TCRF	17.05	13.49	16.06	12.34	19.84	14.59	15.20	13.43	9.56	14.84	14.64
RAP	28.37	16.80	31.05	13.72	17.50	15.69	13.17	15.74	12.38	25.41	18.98
Hier-Joint	30.03	18.74	31.90	15.33	22.91	19.17	15.20	20.04	21.80	32.81	22.79
Hier-Joint$^+$	32.87	19.04	33.54	13.90	24.53	20.72	15.56	23.24	22.65	31.10	23.72
RNSCN	31.41	18.93	31.85	16.73	32.43	25.54	23.31	19.98	19.15	30.56	24.99
RNSCN$^+$	34.60	20.03	35.65	16.59	33.26	26.63	20.04	22.00	18.87	33.21	26.09
SAL	41.64	30.34	42.60	28.00	36.36	32.36	30.14	35.97	26.46	42.18	34.61
AHF	**44.57**	**34.96**	**43.49**	**33.05**	**37.33**	**34.89**	**33.23**	**39.61**	**29.01**	**46.55**	**37.67**

Table 3: Comparison results with baselines.

Model	D → R	D → S	L → R	L → S	R → D	R → L	R → S	S → D	S → L	S → R	Avg
AHF-DD	44.44	33.08	**47.32**	26.60	28.85	31.90	21.13	37.43	**30.68**	40.24	34.17
AHF-AMT	42.74	34.43	40.27	27.60	35.71	33.66	31.86	36.99	26.64	40.84	35.07
AHF-AT+FT1	40.68	28.41	31.02	24.40	34.72	31.33	**33.62**	35.08	27.73	39.54	32.65
AHF-AT+FT2	44.38	28.23	37.54	24.60	36.76	34.77	33.42	36.87	28.99	44.73	35.02
AHF-DW	43.37	33.79	42.85	32.56	37.26	34.45	31.32	38.54	30.09	46.30	37.05
AHF	**44.57**	**34.96**	43.49	**33.05**	**37.33**	**34.89**	33.23	**39.61**	29.01	**46.55**	**37.67**

Table 4: Comparison results for variants of our framework.

- **RNSCN**(Wang and Pan 2018): A recursive neural network for cross-domain aspect and opinion extraction, which relies on the shared dependency structure between different domains.
- **Hier-Joint$^+$/RNSCN$^+$**: Extend Hier-Joint/RNSCN with an additional LSTM layer for the prediction of the aspect boundary labels.
- **SAL**(Li et al. 2019b): A state-of-the-art method that employs dual memory interaction and selective adversarial learning for cross-domain ABSA.

Experimental Results and Analysis

Main Results. The comparison results of our framework AHF with the baselines are presented in Table 3. From the results we find that AHF consistently outperforms the other approaches on all the transfer tasks. Specifically, our approach achieves 3.06% improvements over the state-of-the-art method SAL on average F1 score. Compared with the traditional approaches TCRF and RAP, the neural network based methods achieve better results on most tasks. Hier-Joint and RNSCN leverage the common syntactic dependency relations between different domains to learn the cross-domain features. There always exist some noises in the dependency parsing tree, which will affect their performance. Thus the results of them are lower than SAL which relies on selective adversarial training to align the features. Though SAL achieves better results, the target domain sentences are merely used to guide the feature learning through training the domain discriminator. And our proposed framework AHF can take advantage of the target data for the training of both the domain discriminator and the task classifier. From the results, we can observe that AHF performs better than SAL, which demonstrates the effectiveness of our method.

Ablation Study. To investigate the effect of each component, we perform comparison between the full AHF model and its ablations as shown in Table 4. The variants of our framework includes:

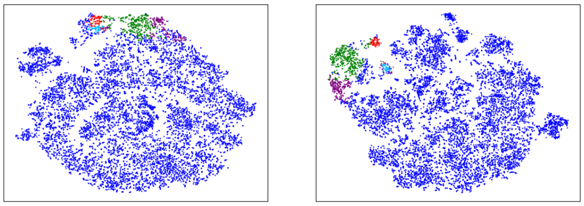
- **AHF-DD**: Remove the domain discriminator from the framework.
- **AHF-AMT**: Combine the student network and the domain discriminator for the domain adaptation.
- **AHF-AT+FT1**: Replace the automatic threshold with a fixed threshold 0.9 as previous work (French, Mackiewicz, and Fisher 2018).
- **AHF-AT+FT2**: Replace the automatic threshold with two fixed thresholds where 0.9 and 0.4 are used for label "O" and aspect labels, respectively.
- **AHF-DW**: Remove the output weight d_i of the domain discriminator from $\mathcal{L}_{se}$.

From the results, we find that the integrated framework performs better than all the variants. The F1 scores of AHF-DD are lower than AHF on most tasks, which demonstrates adversarial training is able to reduce the domain gap. And AHF performs better than AHF-AMT in which the target data are only used to train the domain discriminator. This indicates that the information lying in the target data can be used for the task classifier refinement. The performance will drop when we use the fixed thresholds. Furthermore, we find that the average F1 scores of AHF-AT+FT1 and AHF-AT+FT2 are even lower than AHF-AMT. This proves that using automatic threshold to filter noisy labels is critical to our framework. Comparing the results of AHF and AHF-DW, we can see that the performance of our framework can be slightly improved by the similarity weight.

Feature Visualization. We visualize the feature vectors of the target domain generated by AHF-AMT and AHF. For feature visualization, we adopt t-SNE (Maaten and Hinton 2008) on the S → D task. The results are illustrated in Figure 2, where the "blue" represents non-aspect label "O" and other colors represent aspect labels. The number of different

Input Sentence	SN	AHF-AMT	AHF
1. The included $[\boldsymbol{memory\ card}]_{\boldsymbol{NEG}}$ is too small.	NONE ✗	$[memory\ card]_{NEG}$ ✓	$[memory\ card]_{NEG}$ ✓
2. In all fairness, $[\boldsymbol{customer\ reps}]_{\boldsymbol{POS}}$ are very nice, and they (most of them) try hard to answer your questions.	$[customer\ reps]_{POS}$ ✓	$[customer\ reps]_{NEG}$ ✗	$[customer\ reps]_{POS}$ ✓
3. The $[\boldsymbol{phone\ book}]_{\boldsymbol{POS}}$ is very user-friendly and the $[\boldsymbol{speakerphone}]_{\boldsymbol{POS}}$ is excellent.	$[phone]_{POS}$ ✗ $[speakerphone]_{POS}$ ✓	$[phone]_{POS}$ ✗ $[speakerphone]_{POS}$ ✓	$[phone\ book]_{POS}$ ✓ $[speakerphone]_{POS}$ ✓
4. Lots of flaws, but exceptional $[\boldsymbol{sound\ quality}]_{\boldsymbol{POS}}$, $[\boldsymbol{hd\ size}]_{\boldsymbol{POS}}$, and $[\boldsymbol{price}]_{\boldsymbol{POS}}$ make it a good buy.	$[sound\ quality]_{POS}$ ✓ NONE ✗ NONE ✗	$[sound]_{POS}$ ✗ NONE ✗ NONE ✗	$[sound\ quality]_{POS}$ ✓ $[hd\ size]_{POS}$ ✓ NONE ✗

Table 5: Case study on task S → D. The golden aspect terms and their sentiments are in bold. The "NONE" represents the prediction is empty.



(a) AHF-AMT.

(b) AHF.

Figure 2: Feature visualization with t-SNE algorithm on task S → D.

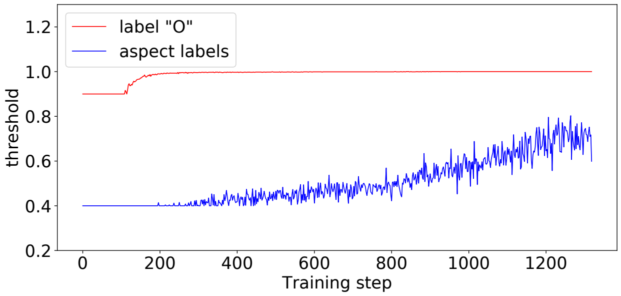


Figure 3: Threshold curves on task S → D.

color points varies greatly due to the data imbalance, where the number of label "O" is much more than others. From the figure, we can observe that there are lots of label ambiguous features generated by AHF-AMT. By contrast, our method produces features that are more easier to distinguish, since it can utilize the target domain aspect and sentiment information to train the task classifier.

Analysis of the Threshold. To illustrate the change of the thresholds for different label sets, we draw the threshold curves with increasing training steps on task S → D. From Figure 3, we find that the thresholds vary considerably for different labels over the training steps. And the threshold of aspect labels is much smaller than that of label "O". The network tends to predict high probability value for label "O" due to the huge number of non-aspect terms in the dataset. The threshold for label "O" is close to 1.0 at about 200 steps, while that value for aspect labels increases gradually during the training steps.

Case Study. We pick some examples from the device dataset and present the predicted results in Table 5. This first column is the input sentence with the golden aspect terms and their sentiment polarities. The other three columns are the prediction results from SN(remove the teacher network and the domain discriminator), AHF-AMT and AHF, respectively. From the first example, we see that AHF-AMT can extract some specific aspects due to the use of adversarial training. But the adversarial loss may trigger a negative transfer, which aligns the target feature with the source feature in an incorrect category. Thus it may fail to predict some aspects that can be identified by SN such as "customer reps" and "sound quality" in the examples. We also find that AHF fails to predict some aspects due to the huge domain discrepancy. But it can predict the aspect terms more accurately than the other two methods. This is because our approach can take advantage of both high quality target domain pseudo labels and adversarial learning.

Conclusion

In this paper, we propose a novel adaptive hybrid framework for cross-domain aspect-based sentiment analysis. We propose to combine the pseudo-label based semi-supervised learning and adversarial learning in a hybrid framework. Thus our model can make use of the target data for both task classifier refinement and domain-invariant features learning. Furthermore, to mitigate the noisy pseudo labels of the target data, we introduce adaptive mean teacher which extends the original Mean Teacher with automatic threshold and domain similarity weight. Extensive experiments on public datasets show that our framework AHF consistently outperforms the state-of-the-art methods.

Acknowledgements

This research is supported in part by the Beijing Municipal Science and Technology Project under Grant Z191100007119008 and the National Natural Science Foundation of China (No. 62006130).

References

Ding, Y.; Yu, J.; and Jiang, J. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *AAAI*.

French, G.; Mackiewicz, M.; and Fisher, M. 2018. Self-ensembling for visual domain adaptation. In *ICLR*, 6.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17(1): 2096–2030.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2018. Adaptive Semi-supervised Learning for Cross-domain Sentiment Classification. In *EMNLP*, 3467–3476.

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2019. An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis. In *ACL*, 504–515.

Hu, M.; and Liu, B. 2004. Mining and summarizing customer reviews. In *KDD*, 168–177.

Jakob, N.; and Gurevych, I. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *EMNLP*, 1035–1045.

Ko, W.-J.; Durrett, G.; and Li, J. J. 2019. Domain agnostic real-valued specificity prediction. In *AAAI*, volume 33, 6610–6617.

Li, F.; Pan, S. J.; Jin, O.; Yang, Q.; and Zhu, X. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *ACL*, 410–419.

Li, X.; Bing, L.; Li, P.; and Lam, W. 2019a. A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. In *AAAI*.

Li, Z.; Li, X.; Wei, Y.; Bing, L.; Zhang, Y.; and Yang, Q. 2019b. Transferable End-to-End Aspect-based Sentiment Analysis with Selective Adversarial Learning. In *EMNLP*, 4582–4592.

Li, Z.; Wei, Y.; Zhang, Y.; and Yang, Q. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *AAAI*.

Li, Z.; Zhang, Y.; Wei, Y.; Wu, Y.; and Yang, Q. 2017. End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification. In *IJCAI*, 2237–2243.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1): 1–167.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL*, 55–60.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*, 43–52.

Mitchell, M.; Aguilar, J.; Wilson, T.; and Van Durme, B. 2013. Open domain targeted sentiment. In *EMNLP*, 1643–1654.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Mohammad, A.-S.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *SemEval 2016*, 19–30.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *SemEval 2015*, 486–495.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *SemEval 2014*, 27–35.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, 1195–1204.

Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; and Kappas, A. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12): 2544–2558.

Toprak, C.; Jakob, N.; and Gurevych, I. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 575–584. ACL.

Wang, W.; and Pan, S. J. 2018. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *ACL*, 2171–2181.

Wang, W.; and Pan, S. J. 2019. Transferable Interactive Memory Network for Domain Adaptation in Fine-Grained Opinion Extraction. In *AAAI*, volume 33, 7192–7199.

Wang, X.; Sun, H.; Qi, Q.; and Wang, J. 2020. SETNet: A Novel Semi-Supervised Approach for Semantic Parsing. In *ECAI*.

Xu, Y.; Du, B.; Zhang, L.; Zhang, Q.; Wang, G.; and Zhang, L. 2019. Self-ensembling attention networks: Addressing domain shift for semantic segmentation. In *AAAI*, volume 33, 5581–5588.

Zhang, K.; Zhang, H.; Liu, Q.; Zhao, H.; Zhu, H.; and Chen, E. 2019. Interactive attention transfer network for cross-domain sentiment classification. In *AAAI*, volume 33, 5773–5780.

Zhang, M.; Zhang, Y.; and Vo, D.-T. 2015. Neural networks for open domain targeted sentiment. In *EMNLP*, 612–621.

Zhou, Y.; Huang, L.; Guo, T.; Han, J.; and Hu, S. 2019. A span-based joint model for opinion target extraction and target sentiment classification. In *IJCAI*, 5485–5491. AAAI Press.