

# Self-training through Classifier Disagreement for Cross-Domain Opinion Target Extraction

Kai Sun\*

sunkai@act.buaa.edu.cn

Beijing Advanced Innovation Center  
for Big Data and Brain Computing,  
School of Computer Science and  
Engineering, Beihang University  
Beijing, China

Nikolaos Aletras

n.aletras@sheffield.ac.uk

Department of Computer Science,  
University of Sheffield,  
Sheffield, United Kingdom

Richong Zhang†

zhangrc@act.buaa.edu.cn

Beijing Advanced Innovation Center  
for Big Data and Brain Computing,  
School of Computer Science and  
Engineering, Beihang University  
Beijing, China

Yongyi Mao

yymao@site.uottawa.ca

School of Electrical Engineering and  
Computer Science, University of  
Ottawa  
Ottawa, Canada

Samuel Mensah\*

s.mensah@sheffield.ac.uk

Department of Computer Science,  
University of Sheffield,  
Sheffield, United Kingdom

Xudong Liu

liuxd@act.buaa.edu.cn

Beijing Advanced Innovation Center  
for Big Data and Brain Computing,  
School of Computer Science and  
Engineering, Beihang University  
Beijing, China

## ABSTRACT

Opinion target extraction (OTE) or aspect extraction (AE) is a fundamental task in opinion mining that aims to extract the targets (or aspects) on which opinions have been expressed. Recent work focus on cross-domain OTE, which is typically encountered in real-world scenarios, where the testing and training distributions differ. Most methods use domain adversarial neural networks that aim to reduce the domain gap between the labelled source and unlabelled target domains to improve target domain performance. However, this approach only aligns feature distributions and does not account for class-wise feature alignment, leading to suboptimal results. Semi-supervised learning (SSL) has been explored as a solution, but is limited by the quality of pseudo-labels generated by the model. Inspired by the theoretical foundations in domain adaptation [2], we propose a new SSL approach that opts for selecting target samples whose model output from a domain-specific teacher and student network disagree on the unlabelled target data, in an effort to boost the target domain performance. Extensive experiments on benchmark cross-domain OTE datasets show that this approach is effective and performs consistently well in settings with large domain shifts.

## CCS CONCEPTS

• **Computing methodologies** → *Information extraction; Semi-supervised learning settings.*

\*Both authors contributed equally to the paper

†Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).  
WWW '23, May 1–5, 2023, Austin, TX, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9416-1/23/04...\$15.00  
<https://doi.org/10.1145/3543507.3583325>

## KEYWORDS

domain adaptation, self-training, opinion mining

## ACM Reference Format:

Kai Sun, Richong Zhang, Samuel Mensah, Nikolaos Aletras, Yongyi Mao, and Xudong Liu. 2023. Self-training through Classifier Disagreement for Cross-Domain Opinion Target Extraction. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*, May 1–5, 2023, Austin, TX, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3543507.3583325>

## 1 INTRODUCTION

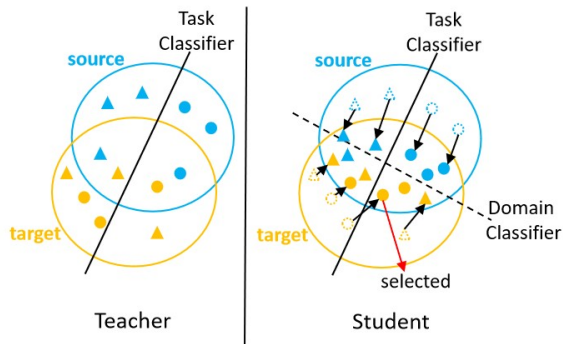
The growth of e-commerce websites has allowed consumers to directly interact with products, leading to an increase in user-generated content. In particular, reviews of products generated by users has grown at an astronomical rate with the increasingly accessibility and affordability of the internet. These reviews, which are often expressed as text, contain sentiment information or opinion words expressed on different aspects of products (referred to as opinion targets or aspect terms). As a result, opinion words along with its corresponding aspect terms have become an important resource to improve recommender systems for Web resource discovery. A typical example is the recommendation of a book in Goodreads [34] based on the opinions expressed on a specific section in the book.

This phenomenon has led to the increased research in opinion mining [25]. This paper focuses on opinion target extraction (OTE) (or aspect extraction (AE)), a fundamental step in opinion mining. AE aims to extract from opinionated sentences the aspects on which opinions have been expressed. Traditional approaches [16, 42] utilize hand-crafted features, which heavily rely on feature extraction. With the advances in deep learning, recent approaches [21, 26, 39] are based on neural networks that are trained in a supervised manner. However, as with any other supervised learning method, these approaches perform poorly when there is a change in domain upon deployment. Cross-domain OTE [9] has emerged as a solution by using unsupervised domain adaptation (UDA) techniques [3, 13, 44]

to reduce the domain shift between a labelled source and unlabelled target domain.

One typical line of work aims to reduce domain shifts via domain adversarial neural networks (DANN) [10]. Given a labelled source and unlabelled target domain data, DANNs attempt to learn representations that are discriminative on the source domain and invariant to the domain distribution. However, DANNs align the feature distributions of the source and target data inputs (i.e., aligning the marginal distribution), neglecting the feature alignment at the class-level [32]. As a consequence, the resulting target features are non-discriminative with respect to the class labels, which consequently leads to suboptimal target domain performance.

Semi-supervised learning (SSL) [4] has been explored to learn target discriminative features by generating pseudo-labels from the unlabelled target data. While SSL approaches have been heavily employed to boost domain adaptation in vision tasks [17, 33, 40], it has been lightly touched in cross-domain OTE [43, 45]. The state-of-the-art method Adaptive Hybrid Framework (AHF) [45] adapts a mean teacher (i.e., teacher and student networks) [33] into the task. The teacher is modelled as a feedforward network while the student is a DANN (i.e., developed by augmenting the feedforward network with a discriminator). Here, knowledge on the target’s output of the teacher-student networks is shared among the networks to learn the target-discriminative features. Although AHF demonstrates the importance of SSL, the fundamental weakness of the mean-teacher cannot be ignored. Specifically, Ke et al. [17] provided theoretical and empirical proof to show that the weights of the teacher quickly converges to that of the student as training progresses, which consequently leads to a performance bottleneck.



**Figure 1: Illustrative example of source and target distributions induced by a teacher and student network (Best viewed in color). Target samples that change class due to adversarial learning by the student network are selected to self-train the student.**

These findings motivate us to decouple the student-teacher networks and optimize the networks through independent paths to prevent the networks from collapsing into each other [17]. We propose a novel SSL approach, which performs Self-training through Classifier Disagreement (SCD), to effectively explore the outputs of the student and teacher networks on the unlabelled target domain. SCD is inspired by the theory of domain adaptation [2], which allows us

to detect high-quality pseudo-labelled target samples in the student feature space to self-train the student for cross-domain OTE. As demonstrated in Fig. 1, SCD achieves this by comparing the two target distributions induced separately by the student and teacher networks. These high-quality pseudo-labelled target samples are those that disagree (i.e., discrepancy in target predictions) with their correspondence in the teacher feature space. We perform extensive experiments and find that SCD not only achieves impressive performance but also performs consistently well in large domain shifts on cross-domain OTE.

Our contribution can be summarized as follows:

- We develop a novel SSL approach for cross-domain OTE, referred to as Self-training through Classifier Disagreement (SCD) which leverages high-quality pseudo-labelled target samples in the student feature space to improve target performance in cross-domain OTE.
- We demonstrate that SCD is favourable in large domain divergence - a key direction in the domain adaptation research.
- We perform extensive experiments and show that SCD achieves state-of-the-art results in nine out of ten transfer pairs for the cross-domain OTE task.

## 2 RELATED WORK

There is a growing literature on OTE [21–23, 26, 39, 41] but they mostly focus on single domain learning. However, in real-world scenarios, the training distribution used by a classifier may differ from the test distribution, which is a big challenge for single domain learning methods.

Cross-domain learning has been explored for the OTE task. Traditional methods use hand-crafted domain-independent features and use Conditional Random Fields (CRFs) [7, 16, 20]. While hand-crafted features are useful, they are manually engineered and require human experts, making them time-consuming and expensive to obtain. So far, some neural models have been proposed for cross-domain OTE [6, 9, 11, 24, 38, 43, 45]. The common paradigm in prior work is to reduce the domain shift between the source and target domains. Among recent work, Ding et al. [9] proposed a hierarchical network trained with joint training (Hier-Joint). This method uses domain-independent rules to generate auxiliary labels and use a recurrent neural network to learn a domain-invariant hidden representation for each word. However, the manually defined rules have limited coverage. A similar method, namely, Recursive Neural Structural Correspondence Network (RNSCN) [38] introduces an opinion word extraction as an auxiliary task based on a critical assumption that associative patterns exist between aspect terms and opinion words irrespective of the domain. They use syntactic relations in dependency trees as the pivot to bridge the domain gap for cross-domain OTE. However, the external linguistic resources used are derived from traditional feature-engineered NLP systems which may propagate errors. More recent methods, including the Aspect Boundary Selective Adversarial Learning model (AD-SAL) [24] uses an adversarial network with attention mechanisms to learn domain-invariant features. Gong et al. [11] proposed BERT<sub>E</sub>-UDA to integrate BERT fine-tuned on domain information for the task.

Chen and Qian [6] proposed a Semantic Bridge network (Sem-Bridge) which constructs syntactic and semantic bridges to transfer common knowledge across domains.

While significant progress has been made, the majority of the proposed models neglect the feature distribution alignment at the class-level. Hence, their performance cannot be guaranteed because they do not learn target discriminative features. Recently, a Cross-Domain Review Generation model based on BERT (BERT<sub>E</sub>-CDRG) [43] generated target domain data with fine-grained annotations aiming to learn the target discriminative features. Perhaps, AHF [45] is the first to use SSL in the task. AHF adapts a mean teacher in which the teacher and student networks are found to be tightly coupled during training, leading to a performance bottleneck [17]. Elsewhere, researchers have delicately designed SSL approaches that allow individual models to iteratively learn from each other, thus, preventing these models from collapsing into each other [5, 17, 31]. Such approaches have demonstrated substantial improvements over the mean-teacher.

### 3 PRELIMINARIES

Our method is inspired by the theory of domain adaptation proposed by Ben-David et al. [2], which provides an upper bound on the target error in terms of the source error and the domain divergence. Suppose  $h \in \mathcal{H}$  is a hypothesis, Ben-David et al. [2] theorized that the target error  $\epsilon_{\mathcal{T}}(h)$  (which can also be viewed as the target performance) is bounded by the source error  $\epsilon_{\mathcal{S}}(h)$  (i.e., the source performance) and the symmetric difference hypothesis divergence  $\mathcal{H}\Delta\mathcal{H}$ -divergence between the source  $\mathcal{S}$  and target  $\mathcal{T}$  distributions, denoted as  $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$  (i.e., a measure of the domain shift). Formally,

$$\forall h \in \mathcal{H}, \epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \beta \quad (1)$$

where  $\beta$  is the optimal joint error on the source and target domains which should be small for domain adaptation. Note,  $\beta$  is a constant which is independent of  $h$ . To obtain a better estimate of  $\epsilon_{\mathcal{T}}(h)$ , a learner can either reduce the source error  $\epsilon_{\mathcal{S}}(h)$  or/and the divergence  $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ , which can be estimated from finite samples of the source and target domains [2].

### 4 PROBLEM STATEMENT

The OTE task is formulated as a sequence labeling problem. Given the  $j$ -th input sentence  $\mathbf{x}_j = \{x_{ij}\}_{i=1}^n$  with  $n$  words, the word  $x_{ij}$  is represented as a feature vector. The goal is to predict the label sequence  $\mathbf{y}_j = \{y_{ij}\}_{i=1}^n$ , with  $y_{ij} \in \mathcal{Y} = \{\text{B, I, O}\}$ , denoting the Beginning, Inside and Outside of an opinion target or aspect term.

In this paper, we focus on the cross-domain setting which is typically tackled through unsupervised domain adaptation (UDA). Particularly, UDA aims to transfer knowledge from a labelled source domain to an unlabelled target domain, whose data distribution has a considerable shift from that of the source domain. Formally, suppose a labelled source domain dataset with  $N_{\mathcal{S}}$  sentence and label pairs  $D_{\mathcal{S}} = \{(\mathbf{x}_j^{\mathcal{S}}, \mathbf{y}_j^{\mathcal{S}})\}_{j=1}^{N_{\mathcal{S}}}$ , and an unlabeled dataset in a target domain with  $N_{\mathcal{T}}$  unlabelled sentences  $D_{\mathcal{T}} = \{(\mathbf{x}_j^{\mathcal{T}})\}_{j=1}^{N_{\mathcal{T}}}$ . Our goal is to

predict labels of testing samples in the target domain using a model trained on  $D_{\mathcal{S}} \cup D_{\mathcal{T}}$ .<sup>1</sup>

## 5 METHODOLOGY

Our method is based on a teacher-student network structure. Teacher  $A$  learns on the source data  $D_{\mathcal{S}}$ ; and Student  $B$  learns on both the source  $D_{\mathcal{S}}$  and target domain data  $D_{\mathcal{T}}$ . Both trained networks generate pseudo-labelled target samples on the unlabelled target domain, which are then compared to detect high quality pseudo-labelled target samples to self-train the student for cross-domain OTE.

### 5.1 Teacher Network

The teacher network  $A = \{A_e, A_l\}$  is a neural network, consisting of a feature encoder  $A_e$  and a label classifier  $A_l$ . In our work,  $A_e$  is modelled using a BiLSTM [14] or BERT [8] since they are both widely used approaches for sequence labelling problems.  $A_l$  on the other hand is modelled using a softmax function. Although the CRF [18] is a typical choice to model the label classifier for sequence labelling problems, the softmax offers comparable performance in cross-domain OTE [24]. Hence, given the sentence  $\mathbf{x}_j = \{x_{ij}\}_{i=1}^n$ ,  $A_e$  extracts the context features  $\mathbf{f}_j^{A_e} = \{f_{ij}^{A_e}\}_{i=1}^n$ . Now, for each word-level feature  $f_{ij}^{A_e}$ , the label classifier  $A_l$  is applied to output the prediction probability  $P(\hat{y}_{ij}^{A_l})$  over the tag set  $\mathcal{Y}$ . As the teacher is trained over the source data only, the classification loss by the teacher network is given by:

$$\mathcal{L}_y^A = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \sum_{i=1}^n \ell(P(\hat{y}_{ij}^{A_l}), y_{ij}) \quad (2)$$

where  $P(\hat{y}_{ij}^{A_l})$  is the probability prediction for the word  $x_{ij} \in \mathbf{x}_j^{\mathcal{S}}$  and  $y_{ij} \in \mathbf{y}_j^{\mathcal{S}}$  is the ground-truth of  $x_{ij}$ .  $\ell$  is the cross-entropy loss function.

Now suppose  $\mathbf{F}_{\mathcal{S}}^{A_e}$  and  $\mathbf{F}_{\mathcal{T}}^{A_e}$  are fixed representations of the respective source and target domain data produced by the trained teacher  $A_e$ . The upper bound on the target error  $\epsilon_{\mathcal{T}}(A_l)$  of the label classifier  $A_l$  can be expressed as:

$$\epsilon_{\mathcal{T}}(A_l) \leq \epsilon_{\mathcal{S}}(A_l) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{F}_{\mathcal{S}}^{A_e}, \mathbf{F}_{\mathcal{T}}^{A_e}) + \beta \quad (3)$$

It is easy to see that the teacher network simply reduces the source error  $\epsilon_{\mathcal{S}}(A_l)$  by (2) while the domain shift  $d_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{F}_{\mathcal{S}}^{A_e}, \mathbf{F}_{\mathcal{T}}^{A_e})$  remains large since the network does not have an appropriate component to reduce the domain shift. This leads to a suboptimal estimate for the bound of the target errors  $\epsilon_{\mathcal{T}}(A_l)$ .

### 5.2 Student Network

As we have seen in the previous section, the teacher applies domain-specific knowledge (i.e., the source domain) for inference, which may underperform on the target domain due to difference in the data distribution. Ideally, the network should have the ability to perform in different domains. We introduce the student network as a solution.

The student network is analogous to a student who learns several subjects simultaneously in order to perform well in those subjects.

<sup>1</sup>Hereinafter, subscripts or superscripts are omitted for clarity, and the term ‘‘aspect’’ will be used instead of ‘‘opinion target’’ to avoid confusion with the target domain.

This is different from teachers who are normally experts in a single subject. This implies that the student network not only desires to be as excellent as the domain-specific teacher on the source data but also aims to perform well on the target data. To this end, the student network is developed by augmenting a teacher network with a discriminator (or domain classifier), following DANN [10]. Accordingly, the student network  $B = \{B_e, B_l, B_d\}$  consists of a feature encoder  $B_e$ ; label classifier  $B_l$ ; and domain classifier  $B_d$ , which determines if the sample comes from the source or target domain.  $B_e$  extracts the context features  $\mathbf{f}_j^{B_e}$  from the sentence  $\mathbf{x}_j \in D_S \cup D_T$  and feeds to  $B_l$  to learn discriminative features on the source domain, following a similar classification loss with Eqn. (2). Formally, the classification loss is defined as:

$$\mathcal{L}_y^B = \frac{1}{N_S} \sum_{j=1}^{N_S} \sum_{i=1}^n \ell(P(\hat{y}_{ij}^{B_l}), y_{ij}) \quad (4)$$

where  $P(\hat{y}_{ij}^{B_l})$  is the probability prediction for the word  $x_{ij} \in \mathbf{x}_j^S$  and  $y_{ij} \in \mathbf{y}_j^S$  is the ground-truth. At the same time,  $\mathbf{f}_j^{B_e}$  is fed to a domain classifier  $B_d$  to learn domain-invariant features through a gradient reversal layer (GRL) [10]. Formally, the GRL  $R_\lambda(\cdot)$  acts as an identity function in the forward pass, i.e.,  $R_\lambda(\mathbf{f}_j^{B_e}) = \mathbf{f}_j^{B_e}$ , and backpropagates the negation of the gradient in the backward pass, i.e.,  $\partial R_\lambda(\mathbf{f}_j^{B_e}) / \partial \mathbf{f}_j^{B_e} = -\lambda \mathbf{I}$ . Consequently,  $B_e$  maximizes the domain classification loss  $\mathcal{L}_d^B$  through the GRL while  $B_d$  minimizes  $\mathcal{L}_d^B$  to make  $\mathbf{f}_j^{B_e}$  domain-invariant. The domain classification loss  $\mathcal{L}_d^B$  is defined as follows:

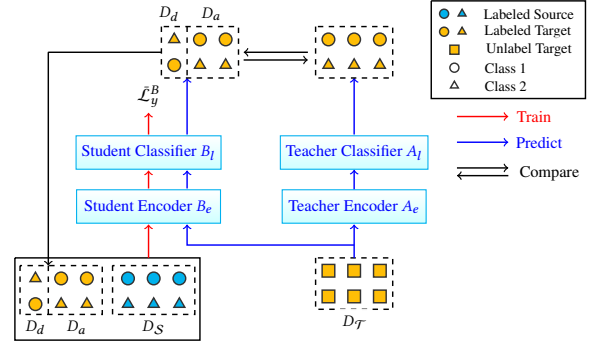
$$\mathcal{L}_d^B = \sum_{j=1}^N d_j \log(P(\hat{d}_j^{B_d})) + (1 - d_j) \log(1 - P(\hat{d}_j^{B_d})) \quad (5)$$

where  $d_j = 1$  indicates that the  $j$ -th sentence comes from the source domain, otherwise  $d_j = 0$ ;  $P(\hat{d}_j^{B_d})$  is the domain probability prediction of the sentence-level feature  $\mathbf{x}_j$ ;  $N = N_S + N_T$ .

Suppose  $\mathbf{F}_S^{B_e}$  and  $\mathbf{F}_T^{B_e}$  are fixed representations of the respective source and target domain data produced by the trained student encoder  $B_e$ . The upper bound on the student label classifier  $B_l$  can be expressed as:

$$\epsilon_{\mathcal{T}}(B_l) \leq \epsilon_S(B_l) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{F}_S^{B_e}, \mathbf{F}_T^{B_e}) + \beta \quad (6)$$

The source error  $\epsilon_S(B_l)$  is comparable with  $\epsilon_S(A_l)$  since the student and teacher are trained on the source data using the same network pipeline (comparing (2) and (4), and also empirically demonstrated in Table 5). But the student network has been shown to reduce the domain divergence with a theoretical guarantee via the GRL [10]. This means  $d_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{F}_S^{B_e}, \mathbf{F}_T^{B_e})$  is relatively small, i.e.,  $d_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{F}_S^{B_e}, \mathbf{F}_T^{B_e}) \leq d_{\mathcal{H}\Delta\mathcal{H}}(\mathbf{F}_S^{A_e}, \mathbf{F}_T^{A_e})$ , and therefore leads to a better estimate of  $\epsilon_{\mathcal{T}}(B_l)$ . In other words, the student performs better than the domain-specific teacher on the target data due to the mitigation of the domain shift.



**Figure 2: Overview of our SSL Approach. Both Teacher and Student networks have been earlier trained by Eqn. (2), (4) and (5). The Student network alone is further self-trained through classifier disagreement on the target domain. This figure is best viewed in color.**

### 5.3 Self-training through Classifier Disagreement

The student network improves target performance by aligning the source and target data distributions. It just so happens that it simply aligns the data distribution without considering the alignment at the class-level [32], leading to suboptimal performance. Such a situation occurs due to the lack of labelled target data to learn target discriminative features. The fundamental challenge is that we do not have access to labelled target data.

To this end, we introduce a strikingly simple approach to collect high-quality pseudo-labelled target samples to improve the class-level alignment of the student network. Fig 2 shows an overview of our approach, which we refer to as Self-training through Classifier Disagreement (SCD). Suppose the trained student and teacher networks (i.e., trained by Eqn. (2), (4) and (5)) assign pseudo-labels to the unlabelled target data. Eqns (3) and (6) indicate that the increase in target performance by the student can be explained by the target samples that have shifted toward the domain-invariant feature space (i.e., the student feature space). Our goal is to self-train the student network by leveraging the target samples responsible for the performance improvement in the target domain.

This strategy is only beneficial if the domain shift is large since this will lead to a large set of high-quality pseudo-labelled target samples. Otherwise, both networks will have comparable performance on the unlabelled target domain and the performance gain is minimal. To extend the approach to problems with close similarity between domains, we split the self-training learning problem by paying attention to: 1)  $D_d$ , the target samples in the student feature space that *disagree* with their counterpart in the teacher feature space; and 2)  $D_a$ , the target samples in the student feature space that *agree* with their counterpart in the teacher feature space.

Formally, let us suppose the student and teacher networks are already trained (i.e., without self-training). As we aim to self-train the Student network, we can rewrite the classification loss expressed in (4) as  $\mathcal{L}_y^{B(0)}$  to represent the initial classification loss of the Student network. Now, let us suppose the teacher and student networks assign the pseudo-labels  $\tilde{\mathbf{y}}_j^{A_l} = \{\tilde{y}_{ij}^{A_l}\}_{i=1}^n$  and  $\tilde{\mathbf{y}}_j^{B_l} = \{\tilde{y}_{ij}^{B_l}\}_{i=1}^n$  for

each sentence  $\mathbf{x}_j^T \in D_{\mathcal{T}}$ , respectively. Self-training is formulated as training the student network on the set  $D_S \cup D_d \cup D_a$ , where the sets  $D_d$  and  $D_a$  are defined as follows:

$$\begin{aligned} D_d &:= \{(\mathbf{x}_j^T, \tilde{\mathbf{y}}_j^{B_l}) | \exists x_{ij} \in \mathbf{x}_j^T \text{ s.t. } \tilde{y}_{ij}^{B_l} \neq \tilde{y}_{ij}^{A_l}\} \\ D_a &:= \{(\mathbf{x}_j^T, \tilde{\mathbf{y}}_j^{B_l}) | \forall x_{ij} \in \mathbf{x}_j^T \text{ s.t. } \tilde{y}_{ij}^{B_l} = \tilde{y}_{ij}^{A_l}\} \end{aligned} \quad (7)$$

Here,  $\tilde{\mathbf{y}}_j^{A_l} \in \tilde{\mathbf{y}}_j^{A_l}$  is the teacher network’s pseudo-label assignment on  $x_{ij} \in \mathbf{x}_j^T$ . Let  $r$  index the self-training round. Then the self-training loss for the student network at a specific self-training round  $r$  can be formulated as follows:

$$\begin{aligned} \mathcal{L}_y^{B(r)} &= \mathcal{L}_y^{B(r)} + \frac{1}{|D_d^{(r)}|} \sum_{(\mathbf{x}_j^T, \tilde{\mathbf{y}}_j^{B_l}) \in D_d^{(r)}} \sum_{x_{ij} \in \mathbf{x}_j^T} \ell(P(\tilde{y}_{ij}^{B_l}), \tilde{y}_{ij}^{B_l}) \\ &\quad + \eta \frac{1}{|D_a^{(r)}|} \sum_{(\mathbf{x}_j^T, \tilde{\mathbf{y}}_j^{B_l}) \in D_a^{(r)}} \sum_{x_{ij} \in \mathbf{x}_j^T} \ell(P(\tilde{y}_{ij}^{B_l}), \tilde{y}_{ij}^{B_l}) \end{aligned} \quad (8)$$

where  $r \geq 1$ ,  $\eta \in [0, 1]$  is a variable to control the weight of the loss on  $D_a^{(r)}$ . Since the similarity between source and target domains can only be measured empirically,  $\eta$  is treated as a hyper-parameter to be tuned.  $\eta$  is expected to be large when the source and target domains are similar, otherwise small. Notice that when  $\eta = 1$ ,  $\tilde{\mathcal{L}}_y^{B(r)}$  becomes a special case of the pseudo-labelling loss function expressed in Eq. 15 of [19] with  $\alpha(t) = 1$ , which we refer to as a standard pseudo-labelling method.

The total loss function  $\mathcal{L}$  for SCD can now be formulated as

$$\mathcal{L} = \mathcal{L}_d^B + \mathcal{L}_y^{B(0)} + \sum_{r \geq 1} \tilde{\mathcal{L}}_y^{B(r)} \quad (9)$$

In each self-training round,  $D_{pl}^{(r)} = D_d^{(r)} \cup D_a^{(r)}$  is generated using the current trained student network. The self-training stops when  $D_{pl}^{(r)}$  is approximately equal in successive rounds.

## 6 EXPERIMENTS AND RESULTS

### 6.1 Experimental Setup

**6.1.1 Comparison Methods.** We evaluate SCD as well as our BERT-based version BERT-SCD in this section. Comparison methods include, CRF [16], FEMA [42], Hier-Joint [9], RN-SCN [38], AD-SAL [24], AHF [45] as well as the BERT-based models BERT<sub>E</sub>-UDA [11] and BERT<sub>E</sub>-CDRG [43]. Two strong single-domain OTE models BERT<sub>B</sub> and BERT<sub>E</sub> [11], which are trained only on the source-domain to investigate the capacity of BERT without domain adaptation. SemBridge [6] is excluded in our comparison since its dataset setup is different from that used in compared works.

**6.1.2 Datasets.** We use benchmark datasets from four domains following previous work [24, 38]. The Laptop dataset consists of reviews in the laptop domain taken from the SemEval ABSA challenge 2014 [30]. The Restaurant dataset is the set of all restaurant reviews in SemEval ABSA challenge 2014, 2015 and 2016 [28–30]. The Device dataset, originally provided by [15] contains reviews in the device domain. The Service dataset, introduced by [35] contains

reviews related to the web service domain. We use the preprocessed data provided by [24]. Dataset statistics are shown in Table 3.

**6.1.3 Evaluation Protocol.** We follow prior work [11, 24] and evaluate on 10 transfer pairs  $D_S \rightarrow D_{\mathcal{T}}$  from the datasets. We use the test set of the source domain as a development set to tune our models. The test set of the target domain is used for evaluation purposes. We evaluate an exact match,<sup>2</sup> and compute the Micro-F1 score. Reported results are the average over 5 runs.

**6.1.4 Implementation Details.** Following Zhou et al. [45], we use 100-dim fixed pretrained Word2Vec embeddings [27] or BERT-Mini embeddings for word features.<sup>3</sup> We use Adam with  $1e^{-3}$  learning rate, 100 epochs for both Teacher and Student networks, and 50 epochs during self-training, word embedding dropout rate in [0.3, 0.5, 0.7], BiLSTM dimensions in [100, 200, 300], adaption rate  $\lambda \in [1.0, 0.7, 0.5, 0.3]$ , batch size in [32, 64, 128] and  $\eta \in [0.0, 0.1, \dots, 0.9, 1.0, 1e^{-2}, 1e^{-3}]$ . Each batch contains half labeled source and half unlabelled target data. All sentences are padded to a max length  $n$ . During self-training, we adopt repeated sampling on the labeled source data with the same size as the pseudo labeled target data in each epoch.

## 6.2 Main Results

Table 1 summarizes our main results. We find that neural methods, including RN-SCN and Hier-Joint surpass hand-crafted feature methods FEMA and CRF, highlighting the importance of leveraging neural networks for the task. We also find that adversarial methods such as AD-SAL and AHF outperforms both Hier-Joint and RN-SCN, indicating that adversarial learning is effective in mitigating the domain shift to yield performance. However, by learning target discriminative features, the SOTA method AHF achieves a better performance over AD-SAL by about 5.45 F1 on average. We see similar performance on the SOTA BERT-based model BERT<sub>E</sub>-CDRG that consider learning target discriminative features. Specifically, BERT<sub>E</sub>-CDRG outperforms the previous SOTA BERT<sub>E</sub>-UDA by about 4.47 F1 on average. This clearly shows the importance of learning target discriminative features. However, AHF considers the a mean teacher while BERT<sub>E</sub>-CDRG considers a generation model to learn these target discriminative features. In contrast, we consider to learn an adversarial model (i.e., Student) based on self-training through classifier disagreement. Our results suggest the effectiveness of our approach where we outperform AHF and BERT<sub>E</sub>-CDRG by an average F1 of 5.92 and 7.08. In particular, we obtain SOTA results on nine out of 10 transfer pairs with relative stability when compared to AHF.

### 6.3 Ablation Study

We study the contribution of model components. Table 2 presents our results. The upper portion of the table shows the performance of different ablated models. The lower portion is the Maximum Mean Discrepancy (MMD) [12], which measures the distance between source and target domain distributions.<sup>4</sup>

<sup>2</sup>Exact Match: the predicted label sequence should exactly match the gold label sequence

<sup>3</sup>We use BERT-Mini implementation from <https://github.com/google-research/bert>

<sup>4</sup>MMD from <https://github.com/easezy/deep-transfer-learning/>

Model	$\mathbb{S} \rightarrow \mathbb{R}$	$\mathbb{L} \rightarrow \mathbb{R}$	$\mathbb{D} \rightarrow \mathbb{R}$	$\mathbb{R} \rightarrow \mathbb{S}$	$\mathbb{L} \rightarrow \mathbb{S}$	$\mathbb{D} \rightarrow \mathbb{S}$	$\mathbb{R} \rightarrow \mathbb{L}$	$\mathbb{S} \rightarrow \mathbb{L}$	$\mathbb{R} \rightarrow \mathbb{D}$	$\mathbb{S} \rightarrow \mathbb{D}$	AVG
CRF	17.00	17.00	2.50	8.80	8.60	4.50	10.90	11.60	9.00	9.70	9.96
FEMA	37.60	35.00	20.70	10.80	14.80	8.80	26.60	15.00	22.90	18.70	21.09
Hier-Joint	52.00	46.70	50.40	19.80	23.40	23.50	31.70	30.00	32.00	33.40	34.29
RNSCN	48.89	52.19	50.39	30.41	31.21	35.50	47.23	34.03	46.16	32.41	40.84
AD-SAL	52.05	56.12	51.55	39.02	38.26	36.11	45.05	35.99	43.76	41.21	43.91
AHF	54.98	58.67	61.11	40.33	47.17	45.78	56.58	36.62	48.24	44.16	49.36 $\pm$ 3.23
SCD	<b>59.52</b>	<b>71.40</b>	<b>61.85</b>	<b>48.30</b>	<b>48.67</b>	<b>52.58</b>	<b>59.68</b>	<b>42.40</b>	<b>54.45</b>	<b>54.01</b>	<b>55.28<math>\pm</math>1.07</b>
BERT <sub>B</sub>	54.29	46.74	44.63	22.31	30.66	33.33	37.02	36.88	32.03	38.06	37.60
BERT <sub>E</sub>	57.56	50.42	45.71	26.50	25.96	30.40	44.18	41.78	35.98	35.13	39.36
BERT <sub>E</sub> -UDA	59.07	55.24	56.40	34.21	30.68	38.25	54.00	44.25	42.40	40.83	45.53
BERT <sub>E</sub> -CDRG	59.17	<b>68.62</b>	58.85	47.61	<b>54.29</b>	42.20	55.56	41.77	35.43	36.53	50.00
BERT-SCD	<b>64.10</b>	67.61	<b>64.75</b>	<b>55.83</b>	51.33	<b>58.92</b>	<b>55.64</b>	<b>49.76</b>	<b>49.62</b>	<b>53.29</b>	<b>57.08<math>\pm</math>1.17</b>

Table 1: Comparison of F1 performance. Best performance is in bold format.

Model	$\mathbb{S} \rightarrow \mathbb{R}$	$\mathbb{L} \rightarrow \mathbb{R}$	$\mathbb{D} \rightarrow \mathbb{R}$	$\mathbb{R} \rightarrow \mathbb{S}$	$\mathbb{L} \rightarrow \mathbb{S}$	$\mathbb{D} \rightarrow \mathbb{S}$	$\mathbb{R} \rightarrow \mathbb{L}$	$\mathbb{S} \rightarrow \mathbb{L}$	$\mathbb{R} \rightarrow \mathbb{D}$	$\mathbb{S} \rightarrow \mathbb{D}$	AVG
SCD	<b>59.52</b>	<b>71.40</b>	<b>61.85</b>	<b>48.30</b>	<b>48.67</b>	<b>52.58</b>	<b>59.68</b>	<b>42.40</b>	<b>54.45</b>	<b>54.01</b>	<b>55.28<math>\pm</math>1.07</b>
SCD( $\eta = 0.0$ )	59.18	<b>71.40</b>	<b>61.85</b>	48.22	48.52	52.25	57.81	40.13	52.78	45.95	53.80 $\pm$ 1.91
SCD( $\eta = 1.0$ )	57.76	67.49	59.06	47.83	46.13	51.03	55.62	<b>42.40</b>	53.80	<b>54.01</b>	53.51 $\pm$ 0.96
Student	55.39	63.69	56.52	47.19	45.48	50.69	52.66	41.22	52.39	44.28	50.95 $\pm$ 1.23
Teacher	52.10	57.46	48.02	24.88	28.48	33.09	48.08	40.92	50.75	45.35	42.87 $\pm$ 1.10
Student(MMD)	0.041	0.040	0.046	0.035	0.094	0.080	0.054	0.042	0.045	0.043	0.052 $\pm$ 0.009
Teacher(MMD)	0.215	0.197	0.415	0.364	0.170	0.263	0.198	0.134	0.158	0.106	0.222 $\pm$ 0.023

Table 2: Ablation Study: F1 Performance of different ablated models (top). Student(MMD) (or Teacher(MMD)) is an estimate of the discrepancy between the learned source and target distributions by the Student (or Teacher).

Dataset	Domain	Sentence	Train	Test
$\mathbb{L}$	Laptop	1869	1458	411
$\mathbb{R}$	Restaurant	3900	2481	1419
$\mathbb{D}$	Device	1437	954	483
$\mathbb{S}$	Service	2153	1433	720

Table 3: Statistics of the datasets.

First, we note that the Teacher and Student networks have comparable performance on the source domain (see results in Table 5). This means the performance of the Student over Teacher is due to the divergence (measured by MMD). Since Student(MMD) is lower than Teacher(MMD) for all transfer pairs, it is not surprising to see the Student network outperforming the Teacher network. Conversely, SCD( $\eta = 1.0$ ) is simply standard pseudo-labelling. Although it improves performance, we find that SCD( $\eta = 0.0$ ) offers comparable performance for the average F1 by focusing on learning only on pseudo-labelled samples with prediction disagreement with the Teacher network. Interestingly, we find that on pairs such as  $\mathbb{S} \rightarrow \mathbb{L}$  and  $\mathbb{S} \rightarrow \mathbb{D}$ , Teacher(MMD) is already low. Although Student(MMD) becomes smaller due to adversarial learning, SCD( $\eta = 0.0$ ) cannot leverage sufficient pseudo-labelled samples to achieve satisfactory performance. This is because Student can only shift few samples to the domain invariant-distribution to bring about a prediction disagreement. But we see the benefit of prediction disagreement on pairs such as  $\mathbb{D} \rightarrow \mathbb{R}$ , where Teacher(MMD) is large

and corresponding Student(MMD) is low, improving the Student network from 56.52 to 61.85 (i.e., performance on SCD( $\eta = 0.0$ )).

These results indicate that the pseudo-labelled samples help to learn the discriminative features, achieving better performance as compared to recent works.

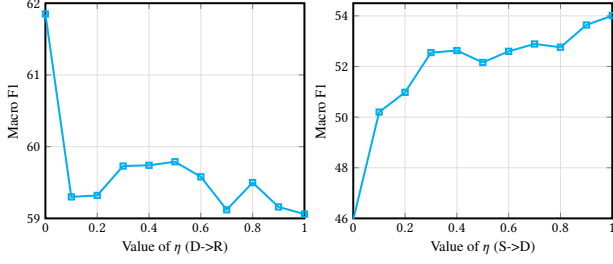
## 6.4 Sensitivity of Hyperparameter $\eta$

We now study the sensitivity of our model for the hyperparameter  $\eta$ . At  $\eta = 0$ , we pay attention to the learning of pseudo-labelled samples by the student network that disagree with those produced by the Teacher network. At  $\eta = 1$ , we are simply performing the standard pseudo-labelling. We study the sensitivity of  $\eta$ , particularly on pairs that have a high or low MMD on the Teacher network. That is, the respective  $\mathbb{D} \rightarrow \mathbb{R}$  and  $\mathbb{S} \rightarrow \mathbb{D}$  pairs. With low MMD, the source and target domains are similar, but diverges with high MMD. The idea is to understand how the domain divergence affects  $\eta$ .

Figure 3 shows the results on this experiment where we report the F1 performance for different values of  $\eta$  on the pairs. We find that on  $\mathbb{S} \rightarrow \mathbb{D}$ , the learning problem moves toward standard pseudo-labelling since the best performance is achieved at  $\eta = 1.0$ . However, on  $\mathbb{D} \rightarrow \mathbb{R}$  the best performance is achieved at  $\eta = 0$ . These results suggest the importance of attention placed on the learning of these pseudo-labelled samples. Particularly, we observe that when the domain divergence is high it is beneficial to learn on pseudo-labelled samples that disagree with the Teacher network. On the other hand, when the source and target domains are similar, pseudo-labelling



seems sufficient for the problem. This model behaviour guides in the selection of  $\eta$ .



**Figure 3: F1 Performance of SCD for different  $\eta$  values on  $\mathcal{D} \rightarrow \mathcal{R}$  (left) and  $\mathcal{S} \rightarrow \mathcal{D}$  (right) .**

### 6.5 Quality of Pseudo-Labels

We perform additional experiments to study the quality of pseudo-labels generated by our method. Figure 5 shows the experiments, where we report the F1 performance for different models for the pairs under study;  $\mathcal{D} \rightarrow \mathcal{R}$  (left) and  $\mathcal{S} \rightarrow \mathcal{D}$  (right). Since SCD( $\eta = 0.0$ ) and SCD have equivalent performance on  $\mathcal{D} \rightarrow \mathcal{R}$  and SCD( $\eta = 1.0$ ) and SCD have equivalent performance on  $\mathcal{S} \rightarrow \mathcal{D}$ , we omit the curves of SCD to clearly show the benefit of pseudo-labelled samples under different strategies. Other compared methods include AHF.

On  $\mathcal{D} \rightarrow \mathcal{R}$ , we find that both SCD( $\eta = 0.0$ ) and SCD( $\eta = 1.0$ ) improves steeply but becomes unstable after the fifth and eighth epochs respectively. However, the improvement of SCD( $\eta = 0.0$ ) over SCD( $\eta = 1.0$ ) is highly notable. This observation points us to the fact, with high Teacher(MMD), prediction disagreement offers high quality pseudo-labelled samples particularly in the early rounds of training to improve performance. However, when Teacher(MMD) is low such as on the  $\mathcal{S} \rightarrow \mathcal{D}$ , we are not able to take advantage of pseudo-labelled samples with prediction disagreement. Hence, the standard pseudo-labelling can outperform prediction disagreement as seen in the figure. AHF on the other hand underperforms, indicating that our SSL approach is effective as compared to the mean teacher.

### 6.6 Feature Visualization

Fig. 4 depicts the t-SNE [36] visualization of features learned using the Teacher, Student and SCD models on the transfer pair  $\mathcal{D} \rightarrow \mathcal{R}$  (1000 instances sampled randomly in each domain). As there are three class labels, namely BIO labels, an ideal model should clearly align the source and target data into three clusters. For the Teacher network, we can observe that the distribution of source samples is relatively far from the distribution of the target samples. Through domain adaptation, the Student network improves the alignment of the source and target samples. However, by learning target discriminative features through SCD, we gradually observe three clusters forming. The results indicate that SCD improves the class-level alignment.

### 6.7 Case Study

To test the effectiveness of our approach, some case examples from the transfer pair with the largest domain divergence ( $\mathcal{D} \rightarrow \mathcal{R}$ ) are

selected for demonstration. Table 4 shows the aspect term extraction results on these case examples.

In the first case, we find that the Teacher, Student and SCD are all capable of identifying the aspect terms “service” and “space”. As these aspect terms appear in both Device and Restaurant domains, domain adaptation is not necessary to extract the aspect terms. It is therefore not surprising to observe that all models identify the aspect terms in the Restaurant domain.

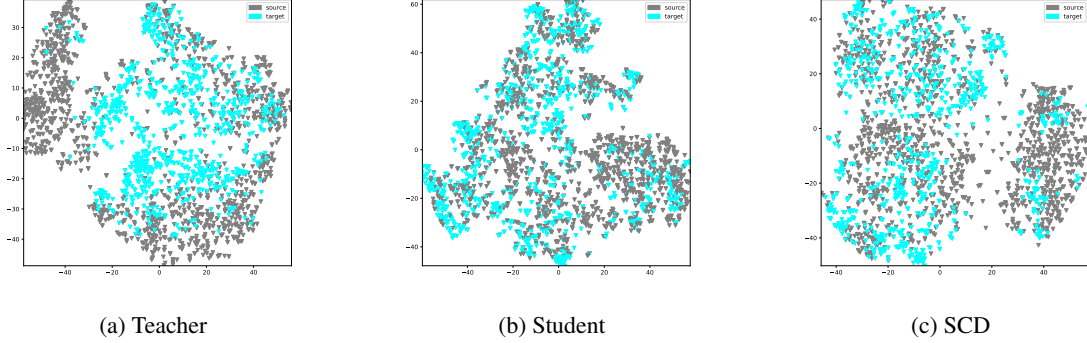
In the second case example, the aspect terms “ambience”, “food” and “catfish” are found in the Restaurant domain and not the Device domain. However, the Teacher was able to extract the aspect terms “ambience” and “food”. Introspecting further, we found that 81% of aspect terms extracted by the Teacher in the Restaurant domain are accompanied with opinion words (e.g., “great”) that are also present in the Device domain. Hence, the Teacher was able to learn the correspondences between opinion words and aspect terms in the Device domain and use that knowledge to locate “ambience” and “food” in the Restaurant domain. However, both Teacher and Student networks fail to extract the aspect term “catfish”. This highlights the importance of learning target discriminative features, as there is no correspondence between the word “delicious” and an aspect term to be learned in the Device domain but only in the Restaurant domain. SCD solves this problem by collecting high quality pseudo-labelled samples in the Restaurant domain. As a result, SCD is able to extract the aspect term “catfish”.

In the third case example, we found that the Teacher network failed to identify the aspect terms “pasta primavera” and “veggies” as they do not exist in the Device domain. However, by reducing the domain shift between the two domains, the Student network is able to extract “pasta primavera” but not “veggies”. Upon investigation, we found that the opinion word “fresh” which expresses an opinion on “veggies” frequently appears 83 times in the Restaurant dataset and 0 times in the Device dataset. Ideally, by learning target discriminative features, we can learn correspondences that exist between “fresh” and aspect terms. Such knowledge as learned by SCD offers supervisory training signals, enabling SCD to detect the aspect term “veggies”.

Finally, in the fourth case example, both the Teacher and Student networks completely failed to detect the aspect term “martinis”. While it is no surprise that the Teacher network fails (i.e., “martinis” is not seen during training), the failure of the Student network highlights the limitations of simply reducing the domain shift and suggests the importance of learning target discriminative features for successful cross-domain OTE.

### 6.8 Performance Comparison on Source Domain

We argued that the difference between the target errors (or F1 performance) of the teacher and student networks can be explained by the  $\mathcal{H}\Delta\mathcal{H}$  divergence when the source errors of these networks are approximately equal. According to Ben-David et al. [2], the source error as well as the divergence can be estimated from finite samples of the source and target domains, under the assumption of the uniform convergence theory [37]. Table 5 therefore reports the F1 performance on the source test set. We discover that for each transfer pair, the F1 performance is approximately equal, comparing the Teacher and Student. This suggest that adversarial learning



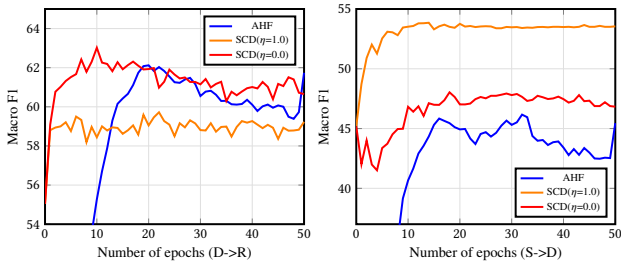
**Figure 4: The t-SNE visualization of features learned by the (a) Teacher, (b) Student, and (c) SCD for the transfer pair  $\mathbb{D} \rightarrow \mathbb{R}$  (light shade: target, dark shade: source)**

Case	Sentence	Teacher	Student	SCD
1	But the <b>space</b> is small and lovely, and the <b>service</b> is helpful.	space, service	space, service	space, service
2	Although small, it has beautiful <b>ambience</b> , excellent <b>food</b> and <b>catfish</b> is delicious.	ambience, food	ambience, food	ambience, food, catfish
3	The <b>pasta primavera</b> was outstanding as well, lots of fresh <b>veggies</b>	NULL	pasta primavera	pasta primavera, veggies
4	I would definitely go back, if only for some of those exotic <b>martinis</b> on the blackboard.	NULL	NULL	martinis

**Table 4: Case study on  $\mathbb{D} \rightarrow \mathbb{R}$ . Gold aspect terms are boldfaced. “NULL” indicates that no aspect term has been extracted.**

Model	$\mathbb{S} \rightarrow \mathbb{R}$	$\mathbb{L} \rightarrow \mathbb{R}$	$\mathbb{D} \rightarrow \mathbb{R}$	$\mathbb{R} \rightarrow \mathbb{S}$	$\mathbb{L} \rightarrow \mathbb{S}$	$\mathbb{D} \rightarrow \mathbb{S}$	$\mathbb{R} \rightarrow \mathbb{L}$	$\mathbb{S} \rightarrow \mathbb{L}$	$\mathbb{R} \rightarrow \mathbb{D}$	$\mathbb{S} \rightarrow \mathbb{D}$	AVG
Teacher	69.63	76.35	66.73	82.67	75.43	67.59	80.00	69.20	79.74	69.16	73.65±0.60
Student	67.48	77.88	65.72	82.63	74.44	67.53	80.17	70.03	80.64	69.41	73.59±0.62

**Table 5: F1 performance of Teacher and Student on the test set of the source domain.**



**Figure 5: F1 performance of different models for training epochs, aiming to evaluate the quality of pseudo labels.**

performed by the Student to reduce the domain shift has little to no effect on the classification on the source data. Most importantly, the results suggest that the difference between the Teacher and Student on the target data is due to the target samples shifted to the domain-invariant space within the Student feature space.

## 7 CONCLUSION

We have proposed a Self-training through Classifier Disagreement for cross-domain OTE. We demonstrated that by simultaneously training a Teacher and a Student network, we can benefit from the information that comes from their predictions on the unlabelled target domain. Specifically, by leveraging pseudo-labelled samples that disagree between the Teacher and Student networks, the Student network is significantly improved, even in large domain divergences. This model behaviour however leads to the potential limitation. In cases of small domain shifts, the model tends to favor pseudo-labeling [19], an SSL approach that risks confirmation bias [33] (i.e., prediction errors are fit by the network). Nevertheless, small domain shifts have little to no interest in cross-domain learning since the source and target domains can be considered to be similar. In the future, we will consider data augmentation strategies to mitigate confirmation bias brought by pseudo-labelling in such situations [1]. We believe our model is generic and can be applied to other cross-domain tasks such as cross-domain named entity recognition.



## ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0110700, in part by the Fundamental Research Funds for the Central Universities, in part by the State Key Laboratory of Software Development Environment. In addition, SM and NA received support from the Leverhulme Trust under Grant Number: RPG#2020#148.

## REFERENCES

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [2] 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* 79, 1 (2010), 151–175.
- [3] Danushka Bollegala, David J. Weir, and John A. Carroll. 2013. Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus. *IEEE Trans. Knowl. Data Eng.* 25, 8 (2013), 1719–1731. <https://doi.org/10.1109/TKDE.2012.103>
- [4] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks* 20, 3 (2009), 542–542.
- [5] Mingcai Chen, Yuntao Du, Yi Zhang, Shuwei Qian, and Chongjun Wang. 2021. Semi-Supervised Learning with Multi-Head Co-Training. *arXiv preprint arXiv:2107.04795* (2021).
- [6] Zhuang Chen and Tiejun Qian. 2021. Bridge-Based Active Domain Adaptation for Aspect Term Extraction. In *ACL/IJCNLP 2021*. 317–327. <https://doi.org/10.18653/v1/2021.acl-long.27>
- [7] Maryna Chernyshevich. 2014. IHS R&D Belarus: Cross-domain extraction of product features using CRF. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*. 309–313. <https://doi.org/10.3115/v1/s14-2051>
- [8] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.
- [9] Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent Neural Networks with Auxiliary Labels for Cross-Domain Opinion Target Extraction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 3436–3442. <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14865>
- [10] 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.* 17 (2016), 59:1–59:35. <http://jmlr.org/papers/v17/15-239.html>
- [11] Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified Feature and Instance Based Domain Adaptation for Aspect-Based Sentiment Analysis. In *EMNLP 2020*. 7035–7045. <https://doi.org/10.18653/v1/2020.emnlp-main.572>
- [12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.
- [13] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Adaptive Semi-supervised Learning for Cross-domain Sentiment Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. 3467–3476. <https://doi.org/10.18653/v1/d18-1383>
- [14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- [15] Mingqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth International Conference on KDD, Seattle, Washington, USA, August 22-25, 2004*. 168–177. <https://doi.org/10.1145/1014052.1014073>
- [16] Niklas Jakob and Iryna Gurevych. 2010. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields. In *EMNLP 2010*. 1035–1045. <https://www.aclweb.org/anthology/D10-1101/>
- [17] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6728–6736.
- [18] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).
- [19] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, Vol. 3.
- [20] Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-Domain Co-Extraction of Sentiment and Topic Lexicons. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*. 410–419. <https://www.aclweb.org/anthology/P12-1043/>
- [21] Kun Li, Chengbo Chen, Xiaojun Qian, Qing Ling, and Yan Song. 2020. Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. 7056–7066. <https://www.aclweb.org/anthology/2020.acl-main.631/>
- [22] Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. 6714–6721. <https://doi.org/10.1609/aaai.v33i01.33016714>
- [23] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect Term Extraction with History Attention and Selective Transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 4194–4200. <https://doi.org/10.24963/ijcai.2018/583>
- [24] Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019. Transferable End-to-End Aspect-based Sentiment Analysis with Selective Adversarial Learning. In *EMNLP-IJCNLP 2019*. 4589–4599. <https://doi.org/10.18653/v1/D19-1466>
- [25] Bing Liu. 2015. *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. Cambridge University Press. <http://www.cambridge.org/us/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/sentiment-analysis-mining-opinions-sentiments-and-emotions>
- [26] Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring Sequence-to-Sequence Learning in Aspect Term Extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 3538–3547. <https://doi.org/10.18653/v1/p19-1344>
- [27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
- [28] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. SemEval-2016 Task 5. In *SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. 19–30. <https://doi.org/10.18653/v1/s16-1002>
- [29] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12. In *SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*. 486–495. <https://doi.org/10.18653/v1/s15-2082>
- [30] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4. In *SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*. 27–35. <https://doi.org/10.3115/v1/s14-2004>
- [31] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*. 135–152.
- [32] Shuhan Tan, Xingchao Peng, and Kate Saenko. 2019. Generalized domain adaptation with covariate and label shift co-alignment. (2019).
- [33] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.. In *ICLR (Workshop)*.
- [34] Mike Thelwall and Kayvan Kousha. 2017. Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology* 68, 4 (2017), 972–983.
- [35] Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. In *ACL 2010*. 575–584. <https://www.aclweb.org/anthology/P10-1059/>
- [36] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [37] Vladimir N Vapnik and A Ya Chervonenkis. 2015. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*. Springer, 11–30.
- [38] Wenya Wang and Sinno Jialin Pan. 2018. Recursive Neural Structural Correspondence Network for Cross-domain Aspect and Opinion Co-Extraction. 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*. 2171–2181. <https://doi.org/10.18653/v1/P18-1202>

- [39] Zhenkai Wei, Yu Hong, Bowei Zou, Meng Cheng, and Jianmin Yao. 2020. Don't Eclipse Your Arts Due to Small Discrepancies: Boundary Repositioning with a Pointer Network for Aspect Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. 3678–3684. <https://www.aclweb.org/anthology/2020.acl-main.339/>
- [40] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10687–10698.
- [41] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. 592–598. <https://doi.org/10.18653/v1/P18-2094>
- [42] Yi Yang and Jacob Eisenstein. 2015. Unsupervised Multi-Domain Adaptation with Feature Embeddings. In *NAACL HLT 2015*. 672–682. <https://doi.org/10.3115/v1/n15-1069>
- [43] Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-Domain Review Generation for Aspect-Based Sentiment Analysis. In *Findings of ACL/IJCNLP 2021*. 4767–4777. <https://doi.org/10.18653/v1/2021.findings-acl.421>
- [44] Guangyou Zhou, Zhiwen Xie, Jimmy Xiangji Huang, and Tingting He. 2016. Bi-Transferring Deep Neural Networks for Domain Adaptation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. <https://doi.org/10.18653/v1/p16-1031>
- [45] Yan Zhou, Fuqing Zhu, Pu Song, Jizhong Han, Tao Guo, and Songlin Hu. 2021. An Adaptive Hybrid Framework for Cross-domain Aspect-based Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14630–14637.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009