Multi-Source Attention for Unsupervised Domain Adaptation

Xia Cui
University of Liverpool
Xia.Cui@liverpool.ac.uk

Danushka Bollegala*
University of Liverpool, Amazon
danushka@liverpool.ac.uk

Abstract

We model source-selection in multi-source Unsupervised Domain Adaptation (UDA) as an attention-learning problem, where we learn attention over the sources per given target instance. We first independently learn source-specific classification models, and a relatedness map between sources and target domains using pseudo-labelled target domain instances. Next, we learn domain-attention scores over the sources for aggregating the predictions of the source-specific models. Experimental results on two cross-domain sentiment classification datasets show that the proposed method reports consistently good performance across domains, and at times outperforming more complex prior proposals. Moreover, the computed domain-attention scores enable us to find explanations for the predictions made by the proposed method. [1]

1 Introduction

Domain adaptation (DA) considers the problem of generalising a model learnt using the data from a particular source domain to a different target domain (Zhang et al., 2015). Although most DA methods consider adapting to a target domain from a single source domain (Blitzer et al., 2006, 2007; Ganin et al., 2016), often it is difficult to find a suitable single source to adapt from, and one must consider multiple sources. For example, in sentiment classification, each product category is considered as a *domain* (Blitzer et al., 2006), resulting in a multi-domain adaptation setting.

Unsupervised DA (UDA) is a special case of DA where labelled instances are not available for the target domain. Existing approaches for UDA can be categorised into pivot-based and instance-based methods. Pivots refer to the features common to both source and target domains (Blitzer et al., 2006). Pivot-based single-source domain adaptation methods, such as Structural Correspondence Learning (SCL; Blitzer et al., 2006, 2007) and Spectral Feature Alignment (SFA; Pan et al., 2010), first select a set of pivots and then project the source and target domain documents into a shared space. Next, a prediction model is learnt in this shared space. However, these methods fail in multi-source settings because it is challenging to find pivots across all sources such that a single shared projection can be learnt. Similarly, instance-based methods, such as Stacked Denoising Autoencoders (SDA; Glorot et al., 2011) and marginalised SDA (mSDA; Chen et al., 2012) minimise the loss between the original inputs and their reconstructions. Not all of the source domains are appropriate for learning transferable projections for a particular target domain. Adapting from an unrelated source can result in poor performance on the given target, which is known as *negative transfer* (Rosenstein et al., 2005; Pan and Yang, 2010; Guo et al., 2018).

Prior proposals for multi-source UDA can be broadly classified into methods that: (a) first select a source domain and then select instances from that source domain to adapt to a given target domain test instance (Ganin et al., 2016; Kim et al., 2017; Zhao et al., 2018; Guo et al., 2018); (b) pool all source domain instances together and from this pool select instances to adapt to a given target domain test instance (Chattopadhyay et al., 2012); (c) pick a source domain and use all instances in that source (source domain selection) (Schultz et al., 2018); and (d) pick all source domains and use all instances (utilising all instances) (Aue and Gamon, 2005; Bollegala et al., 2011; Wu and Huang, 2016).

In contrast, we propose a multi-source UDA

*Danushka Bollegala holds concurrent appointments as a Professor at University of Liverpool and as an Amazon Scholar. This paper describes work performed at the University of Liverpool and is not associated with Amazon.

[1] Source code available at https://github.com/LivNLP/multi-source-attention

method that systematically addresses the various challenges in multi-source UDA.

- Although in UDA we have labelled instances in each source domain, its number is significantly smaller than that of the unlabelled instances in the same domain. For example, in the Amazon product review dataset released by Blitzer et al. (2007) there are 73679 unlabelled instances in total across the four domains, whereas there are only 4800 labelled instances. To increase the labelled instances in a source domain, we induce pseudo-labels for the unlabelled instances in each source domain using self-training as in § 3.1.
- In UDA, we have no labelled data for the target domain. To address this challenge, we infer pseudo-labels for the target domain's unlabelled training instances by majority voting over the classifiers trained from each source domain, using both labelled and pseudo-labelled instances as in § 3.1.
- Given that the pseudo-labels inferred for the target domain instances are inherently more noisier compared to the manually labelled source domain instances, we propose a method to identify a subset of prototypical target domain instances for DA using document embedding similarities as described in § 3.2.
- The accuracy of UDA is upper-bounded by the $\mathcal{H}$-divergence between a source and a target domain (Kifer et al., 2004; Ben-David et al., 2006, 2009). Therefore, when predicting the label of a target domain test instance, we must select only the relevant labelled instances from a source domain. We propose a method to learn such a *relatedness map* between source and target domains in § 3.3.
- To reduce negative transfer, for each target domain test instance we dynamically compute a *domain-attention* score that expresses the relevance of a source domain. For this purpose, we represent each domain by a *domain embedding*, which we learn in an end-to-end fashion using the target domain's pseudo-labelled instances as detailed in § 3.4.

We evaluate the proposed method on two standard cross-domain sentiment classification benchmarks for UDA. We find that both pseudo-labels and domain-attention scores contribute toward improving the classification accuracy for a target domain. The proposed method reports consistently good performance in both datasets and across multiple domains. Although the proposed method does not outperform more complex UDA methods in some cases, using the domain-attention scores, we are able to retrieve justifications for the predicted labels.

2 Related Work

In § 1 we already mentioned prior proposals for single-source DA and this section discusses multi-source DA, which is the main focus of this paper. Bollegala et al. (2011) created a sentiment sensitive thesaurus (SST) using the data from the union of multiple source domains to train a cross-domain sentiment classifier. The SST is used to expand feature spaces during train and test times. The performance of SST depends heavily on the selection of pivots (Cui et al., 2017; Li et al., 2017). Wu and Huang (2016) proposed a sentiment DA method from multiple sources (SDAMS) by introducing two components: a sentiment graph and a domain similarity measure. The sentiment graph is extracted from unlabelled data. Similar to SST, SDAMS uses data from multiple sources to maximise the available labelled data. Guo et al. (2020) proposed a mixture of distance measures and used a multi-arm bandit to dynamically select a single source during training. However, in our proposed method all domains are selected and contributing differently as specified by their domain-attention weights for each train and test instance. Moreover, we use only one distance measure and is easier to implement.

Recently, Adversarial NNs have become popular in DA (Ganin et al., 2016; Zhao et al., 2018; Guo et al., 2018). Adversarial training is used to reduce the discrepancy between source and target domains (Ding et al., 2019). Domain-Adversarial Neural Networks (DANN; Ganin et al., 2016) use a gradient reversal layer to learn domain independent features for a given task. Multiple Source Domain Adaptation with Adversarial Learning (MDAN; Zhao et al., 2018) generalises DANN and aims to learn domain independent features while being relevant to the target task. Li et al. (2017) proposed End-to-End Adversarial Memory Network (AMN), inspired by memory networks (Sukhbaatar et al., 2015), and automatically capture pivots using an attention mechanism. Guo et al. (2018) proposed an UDA method using a mixture of experts for each domain. They model the domain relations using a *point-to-set* distance metric to the encoded training matrix for source domains. Next, they perform joint training over all domain-pairs to update the parameters in the model by *meta-training*. However, they ignore the available unlabelled instances for the source domain. Adversarial training methods have shown to be sensitive to the hyper parameter values and require problem-specific techniques (Mukherjee et al., 2018). Kim et al. (2017) modeled domain relations using *example-to-domain* based on an attention mechanism. However, the attention weights are learnt using source domain training data in a supervised manner. Following a self-training approach, Chattopadhyay et al. (2012) proposed a two-stage weighting framework for multi-source DA that first computes the weights for features from different source domains using Maximum Mean Discrepancy (MMD; Borgwardt et al., 2006). Next, they generate pseudo labels for the target unlabelled instances using a classifier learnt from the multiple source domains. Finally, a classifier is trained on the pseudo-labelled instances for the target domain. Their method requires labelled data for the target domain, which is a *supervised* DA setting, different from the UDA setting we consider in this paper. Our proposed method uses self-training to assign pseudo-labels for the unlabelled target instances, and learn an embedding for each domain using an attention mechanism.

3 Multi-Source Domain Attention

Let us assume that we are given N source domains, $S_1, S_2, \ldots, S_N$, and required to adapt to a target domain T. Moreover, let us denote the labelled instances in S_i by $\mathcal{S}_i^L$ and unlabelled instances by $\mathcal{S}_i^U$. For T we have only unlabelled instances $\mathcal{T}^U$ in UDA. Our goal is to learn a binary classifier[2] to predict labels ($\in \{0, 1\}$) for the target domain instances using $\mathcal{S}^L = \cup_{i=1}^N \mathcal{S}_i^L$, $\mathcal{S}^U = \cup_{i=1}^N \mathcal{S}_i^U$ and $\mathcal{T}^U$. We denote labelled and unlabelled instances in S_i by respectively x_i^L and x_i^U, whereas instances in T are denoted by x_T. To simplify the notation, we drop the superscripts L and U when it is clear from the context whether the instance is respectively labelled or not.

The steps of our proposed method can be summarised as follows: (a) use labelled and unlabelled instances from each of the source domains to learn classifiers that can predict the label for a given instance. Next, develop a majority voter and use it to predict the *pseudo-labels* for the target domain unlabelled instances $\mathcal{T}^U$ (§ 3.1); (b) compute a *relatedness map* between the target domain's pseudo-labelled instances, $\mathcal{T}^{L*}$, and source domains' labelled instances $\mathcal{S}^L$ (§ 3.3); (c) compute *domain-attention* weights for each source domain (§ 3.4); (d) jointly learn a model based on the relatedness map and the domain-attention weights for predicting labels for the target domain's test instances (§ 3.5).

[2] Although we consider binary sentiment classification as an evaluation task in this paper, the proposed method can be easily extended to multi-class classification settings by making 1-vs-rest prediction tasks (Rifkin and Klautau, 2004).

3.1 Pseudo-Label Generation

In UDA, we have only unlabelled data for the target domain. Therefore, we first infer pseudo-labels for the target domain instances $\mathcal{T}^U$ by self-training (Abney, 2007) following Algorithm 1. Specifically, we first train a predictor f_i for the i-th source domain using only its labelled instances $\mathcal{S}_i^L$ using a base learner Γ (Line 1-2). Any classification algorithm that can learn a predictor f_i that can compute the probability, $f_i(x, y)$, of a given instance x belonging to the class y can be used as Γ. In our experiments, we use logistic regression for its simplicity and popularity in prior UDA work (Bollegala et al., 2011; Bollegala et al., 2013).

Next, for each unlabelled instance in the selected source domain, we compute the probability of it belonging to each class and find the most probable class label. If the probability of the most likely class is greater than the given confidence threshold $\tau \in [0, 1]$, we will append that instance to the current labelled training set. This enables us to increase the labelled instances for the source domains, which is important for learning accurate classifiers when the amount of labelled instances available is small. After processing all unlabelled instances in S_i, we train the final classifier f_i for S_i using both original and pseudo-labelled instances. Finally, we predict a pseudo-label for a target domain instance as the majority vote, $f^* \in \{0, 1\}$, over the predictions made by the individual classifiers f_i.

3.2 Prototype Selection

Selecting the highest confident pseudo-labelled instances for training a classifier for the target domain as done in prior work (Zhou and Li, 2005; Abney, 2007; Søgaard, 2010; Ruder and Plank, 2018) does not guarantee that those instances will be the most

Algorithm 1 Multi-Source Self-Training

Input: source domains' labelled instances $\mathcal{S}_1^L, \ldots, \mathcal{S}_N^L$, source domains' unlabelled instances $\mathcal{S}_1^U, \ldots, \mathcal{S}_N^U$ and target domain's unlabelled instances $\mathcal{T}^U$, target classes $\mathcal{Y}$, base learner Γ and the classification confidence threshold τ.

Output: multi-source self-training classifier f^*

for $i = 1$ **to** N **do**
  $\mathcal{L}_i \leftarrow \mathcal{S}_i^L$
  $f_i \leftarrow \Gamma(\mathcal{L}_i)$
  for $x \in \mathcal{S}_i^U$ **do**
    $\hat{y} = \arg\max_{y \in \mathcal{Y}} f_i(x, y)$
    if $f_i(x, \hat{y}) > \tau$ **then**
      $\mathcal{L}_i \leftarrow \mathcal{L}_i \cup \{(x, \hat{y})\}$
    end if
  end for
  $f_i \leftarrow \Gamma(\mathcal{L}_i)$
end for
return majority voter f^* over $f_1, \ldots, f_N$.

suitable ones for adapting to the target domain, which was not considered during the self-training stage. For example, some target instances might not be good prototypical examples of the target domain and we would not want to use the pseudo-labels induced for those instances when training a classifier for the target domain. To identify instances in the target domain that are better prototypes, we first encode each target instance by a vector and select the instances that are closest to the centroid, $\boldsymbol{c}_T$, of the target domain instances given by (1).

$$\boldsymbol{c}_T = \frac{1}{|\mathcal{T}^U|} \sum_{x \in \mathcal{T}^U} \boldsymbol{x} \tag{1}$$

In the case of text documents x, their embeddings, $\boldsymbol{x}$, can be computed using numerous approaches such as using bi-directional LSTMs (Melamud et al., 2016) or transformers (Reimers and Gurevych, 2019). In our experiments, we use the Smoothed Inversed Frequency (SIF; Arora et al., 2017), which computes document embeddings as the weighted-average of the pre-trained word embeddings for the words in a document. Despite being unsupervised, SIF has shown strong performance in numerous semantic textual similarity benchmarks (Agirre et al., 2015). Using the centroid computed in (1), similarity for target instance to the centroid is computed using the cosine similarity given in (2).

$$\text{sim}(\boldsymbol{x}, \boldsymbol{c}_T) = \frac{\boldsymbol{x}^\top \boldsymbol{c}_T}{||\boldsymbol{x}||\,||\boldsymbol{c}_T||} \tag{2}$$

Other distance measures such as the Euclidean distance can also be used. We use cosine similarity here for its simplicity. We predict the labels for the target domain unlabelled instances, $\mathcal{T}^U$, using f^*, and select the instances with the top-k highest similarities to the target domain according to (2) as the target domain's pseudo-labelled instances $\mathcal{T}^{L*}$.

3.3 Relatedness Map Learning

Not all of the source domain instances are relevant to a given target domain instance and the performance of a classifier under domain shift can be upper bounded by the $\mathcal{H}$-divergence between a source and a target domain (Kifer et al., 2004; Ben-David et al., 2006, 2009). To model the relatedness between a target domain instance and each instance from the N source domains, we use the pseudo-labelled target domain instances $\mathcal{T}^{L*}$ and source domains' labelled instances $\mathcal{S}_i^L$ to learn a *relatedness map*, ψ_i, between a target domain instance $\boldsymbol{x}_T (\in \mathcal{T}^{L*})$ and a source domain labelled instance $\boldsymbol{x}_i^L$ $(\in \mathcal{S}_i^L)$ as given by (3).

$$\psi_i(\boldsymbol{x}_T, \boldsymbol{x}_i^L) = \frac{\exp({\boldsymbol{x}_T}^\top \boldsymbol{x}_i^L)}{\sum_{\boldsymbol{x}' \in \mathcal{S}_i^L} \exp({\boldsymbol{x}_T}^\top \boldsymbol{x}')} \tag{3}$$

Using ψ_i, we can determine how well each instance in a source domain contributes to the prediction of the label of a target domain's instance.

3.4 Instance-based Domain-Attention

To avoid negative transfer, we dynamically select the source domain(s) to use when predicting the label for a given target domain instance. Specifically, we learn *domain-attention*, $\theta(\boldsymbol{x}_T, \mathcal{S}_i)$, for each source domain, S_i, conditioned on x_T as given by (4).

$$\theta(\boldsymbol{x}_T, \mathcal{S}_i) = \frac{\exp({\boldsymbol{x}_T}^\top \boldsymbol{\phi}_i)}{\sum_{j=1}^{N} \exp({\boldsymbol{x}_T}^\top \boldsymbol{\phi}_j)} \tag{4}$$

$\boldsymbol{\phi}_i$ can be considered as a *domain embedding* for S_i and has the same dimensionality as the instance embeddings. During training we initialise $\boldsymbol{\phi}_i$ using Xavier initialisation (Glorot and Bengio, 2010) and normalise such that $\forall \boldsymbol{x}_T, \sum_{i=1}^{N} \theta(\boldsymbol{x}_T, \mathcal{S}_i) = 1$.

3.5 Training

We combine the relatedness map (§ 3.3) and domain-attention (§ 3.4) and predict the label, $\hat{y}(x_T)$, of a target domain instance x_T using (5).

$$\hat{y}(x_T) = \sigma \left(\sum_{i=1}^{N} \sum_{\boldsymbol{x}_i^L \in \mathcal{S}_i^L} y(x_i^L)\, \psi_i(\boldsymbol{x}_T, \boldsymbol{x}_i^L)\, \theta(\boldsymbol{x}_T, \mathcal{S}_i) \right) \quad (5)$$

Here, $\sigma(z) = 1/(1 + \exp(-z))$ is the logistic sigmoid function and $y(x_i^L)$ is the label of the source domain labelled instance x_i^L. First, we use the target instances, $x \in \mathcal{T}^{L*}$, with inferred labels $y^*(x)$ (computed using f^* from Algorithm 1) as the training instances and predict their labels, $\hat{y}(x)$, by (5). The cross entropy error, $E(\hat{y}(x), y^*(x))$ for this prediction is given by (6):

$$E(\hat{y}(x), y^*(x)) = -\lambda(x)(1-y^*(x)) \log(1-\hat{y}(x)) - \lambda(x) y^*(x) \log(\hat{y}(x)) \quad (6)$$

Here, $\lambda(x)$ is a rescaling factor computed using the normalised similarity score as in (7):

$$\lambda(x) = \frac{\mathrm{sim}(\boldsymbol{x}, \boldsymbol{c}_T)}{\sum_{\boldsymbol{x}' \in \mathcal{T}^{L*}} \mathrm{sim}(\boldsymbol{x}', \boldsymbol{c}_T)} \quad (7)$$

We minimise (6) using ADAM (Kingma and Ba, 2015) for learning the domain-embeddings, $\boldsymbol{\phi}_i$. The initial learning rate is set to 10^{-3} using a subset of $\mathcal{T}^{L*}$ held-out as a validation dataset.

4 Experiments

To evaluate the proposed method, we use the multi-domain Amazon product review dataset compiled by Blitzer et al. (2007). This dataset contains product reviews from four domains: Books (**B**), DVD (**D**), Electronics (**E**) and Kitchen Appliances (**K**). Following Guo et al. (2018), we conduct experiments under two different splits of this dataset as originally proposed by Blitzer et al. (2007) (**Blitzer2007**) and by Chen et al. (2012) (**Chen2012**). Table 1 shows the number of instances in each dataset. By using these two versions of the Amazon review dataset, we can directly compare the proposed method against relevant prior work. Next, we describe how the proposed method was trained on each dataset.

For **Blitzer2007**, we use the official train and test splits where each domain contains 1600 labelled training instances (800 positive and 800 negative), and 400 target test instances (200 positive and 200 negative). In addition, each domain also contains 6K-35K unlabelled instances. We use 300 dimensional pre-trained GloVe embeddings (Pennington et al., 2014) following prior work (Bollegala et al., 2011; Wu and Huang, 2016) with SIF to create document embeddings for the reviews.

In **Chen2012**, each domain contains 2000 labelled training instances (1000 positive and 1000 negative), and 2000 target test instances (1000 positive and 1000 negative). The remainder of the instances are used as unlabelled instances (ca. 4K-6K for each domain). We use the publicly available[3] 5000 dimensional tf-idf vectors produced by Zhao et al. (2018). We use a multilayer perceptron (MLP) with an input layer of 5000 dimensions and 3 hidden layers with 500 dimensions. We use final output layer with 500 dimensions as the representation of an instance.

For each setting, we follow the standard input representation methods as used in prior work. It also shows the flexibility of the proposed method to use different (embedding vs. BoW) text representation methods. We conduct experiments for cross-domain sentiment classification with multiple sources by selecting one domain as the target and the remaining three as sources. The statistics for the two settings are shown in Table 1.

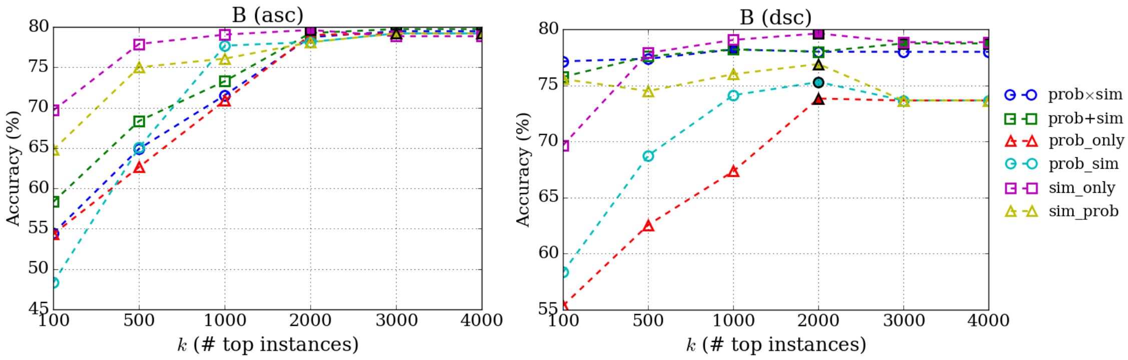
4.1 Effect of Self-Training

As described in § 3.1, our proposed method uses self-training to generate pseudo-labels for the target domain unlabelled instances. In Table 2, we compare self-training against alternative pseudo-labelling methods on **Chen2012**: Self-Training (**Self**; Abney, 2007; Chattopadhyay et al., 2012), Union Self-Training (**uni-Self**; Aue and Gamon, 2005), Tri-Training (**Tri**; Zhou and Li, 2005) and Tri-Training with Disagreement (**Tri-D**; Søgaard, 2010). We observe that all semi-supervised learning methods improve only slightly over **uni-MS**, the baseline model trained on the union of all sources and tested directly on a target domain without any DA, which has been identified as a strong baseline for multi-source DA (Aue and Gamon, 2005; Zhao et al., 2018; Guo et al., 2018). Therefore, pseudo-labelling step alone is insufficient for DA. Moreover, we observe that all semi-supervised methods perform comparably.

[3] https://github.com/KeiraZhao/MDAN/

Target	Source	Train	Test	Unlabel	Train	Test	Unlabel
		Blitzer2007 (Blitzer et al., 2006)			Chen2012 (Chen et al., 2012)		
B	D,E,K	1600×3	400	6000	2000×3	2000	4465
D	B,E,K	1600×3	400	34741	2000×3	2000	5586
E	B,D,K	1600×3	400	13153	2000×3	2000	5681
K	B,D,E	1600×3	400	16785	2000×3	2000	5945

Table 1: Number of train, test and unlabelled instances for the two Amazon product review datasets.

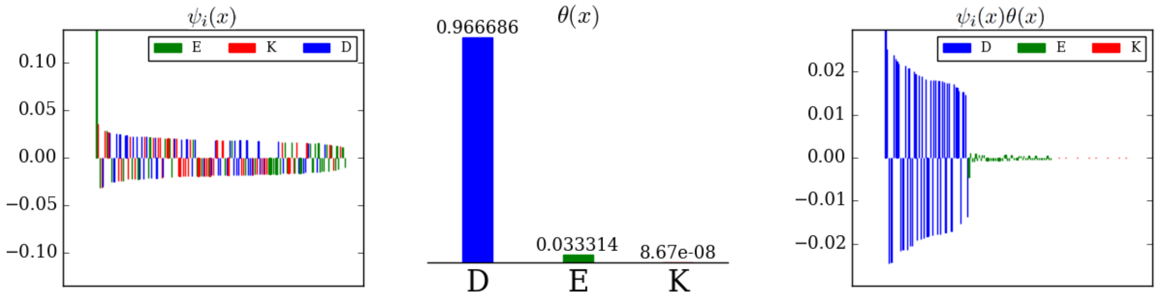


(a) prob sorted in ascending order

(b) prob sorted in descending order

Figure 1: The number of selected pseudo-labelled instances k on **Blitzer2007** is shown on the x-axis. prob denotes prediction confidence from the pseudo classifier trained on the source domains, sim denotes the similarity to the target domain, asc and dsc respectively denote sorted in ascending and descending order (only applied to prob related selection methods, sim is always sorted in dsc). prob_only denotes using only prediction confidence, sim_only denotes using only target similarity. prob_sim indicates selecting by prob first and then sim (likewise for sim_prob). prob×sim denotes using the product of prob and sim, and prob+sim denotes using their sum. The marker for the best result of each method is filled.

Example (1) *Why anybody everest feet would want reading this? ... pure pleasure why 29028 feet account this?... It's a pleasure to read.*



(a)

(b)

(c)

Figure 2: A positively labelled a target test instance in **B** (top) and resulted θ, ψ_i and the product of ψ_i and θ (bottom). Here, the x-axis represents the instances and the y-axis represents the prediction scores. Instance specific values in (a) and (c) are shown as > 0 for positive labelled instances and otherwise < 0. Source instances from **D**, **E** and **K** are shown in blue, green and red respectively. The contributions from top-150 instances from three source domains are shown.

4.2 Pseudo-labelled Instances Selection

When selecting the pseudo-labelled instances from the target domain for training a classifier for the target domain, we have two complementary strategies: (a) select the most confident instances according to f^* (denoted by *prob*) or (b) select the most similar instances to the target domain's centroid (denoted by *sim*). To evaluate the effect of these two strategies and their combinations (i.e prob+sim and prob×sim), in Figure 1, we select target instances with each strategy and measure the accuracy on the target domain **B** for increasing numbers of in-

Example (2) *Her relationship limited own pass her own analysis, there're issues mainly focus in turn for codependency. Disappointing, dysfunctional. Mother'll book her daughter's turn the pass, message turn the message issues analysis of very disappointing information.*

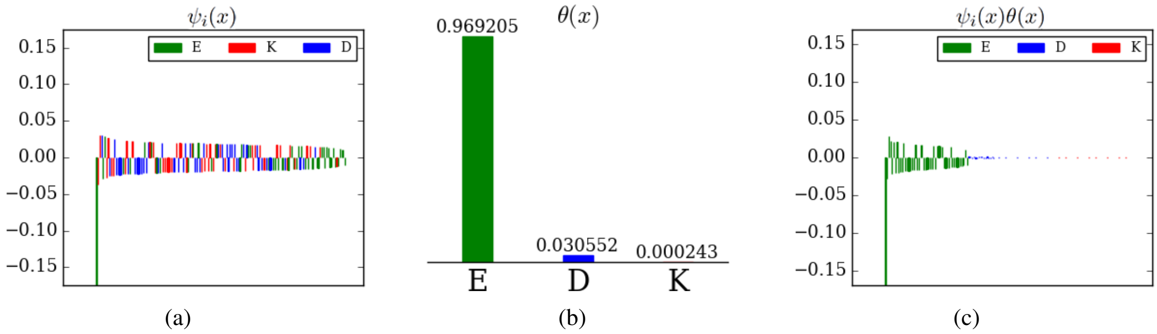


Figure 3: A negatively labelled target test instance in **B**.

DM	L	Score	Evidences (Reviews)
E	-	0.16943	Serious problems.
E	-	0.02823	Sound great but lacking isolation in other areas.
E	+	0.02801	Cases for the cats walking years, no around and knocking...walking on similar cases of cats.
E	+	0.02233	Cord supposed to no problems, this extension extension not worked as cord did...whatever expected just worked fine.
E	-	0.02209	Buy this like characters not used names...be aware of many commonly used characters before you accept file like drive.

Table 3: The top-5 evidences for Example (2) selected from the source domains. DM denotes the domain of the instance. L denotes the label for the instance. Score is $\psi_i(x)\theta(x)$.

T	uni-MS	Self	uni-Self	Tri	Tri-D
B	79.46	79.60	79.46	**79.61**	79.51
D	82.32	**82.49**	82.35	82.35	82.35
E	84.93	84.97	84.93	**84.99**	84.93
K	87.17	87.18	87.17	87.15	**87.23**

Table 2: Classification accuracies (%) for semi-supervised methods on **Chen2012**.

stances k in the descending (dsc) and ascending (asc) order of the selection scores.

From Figure 1b we observe that selecting the highest confident instances does not produce the best UDA accuracies. In fact, merely selecting instances based on confidence scores only (corresponds to prob_only) reports the worst performance. On the other hand, instances that are highly similar to the target domain's centroid are more effective for DA. We observe that with only $k = 1000$ instances, sim_only reaches almost its optimal accuracy. Using validation data, we estimated that $k = 2000$ to be sufficient for all domains to reach the peak performance regardless of the selection strategy. Therefore, we selected 2000 pseudo-labelled instances for the attention step. In our experiments, we used sim_only to select pseudo-labelled instances because it steadily improves the classification accuracy with k for all target domains, and is competitive against other methods.

4.3 Effect of the Relatedness Map

In Table 4 we report the classification accuracy on the test instances in the target domain over the different steps: **uni-MS** (no adapt baseline), **Self** (self-training), **PL** (pseudo-labelling) and **Att** (attention). We use the self-training method described in Algorithm 1. The results clearly demonstrate a consistent improvement over all the steps in the proposed method. For **Self** step, the proposed method improves the accuracy only slightly without any information from the target domain. In the **PL** step, we report the results of a predictor trained on target pseudo-labelled instances. We report the evaluation results for the trained attention model in **Att**.

In **Att** step, we use the relatedness map ψ_i to express the similarity between a target instance and each of source domain instances, and the domain attention score θ to express the relation between a target instance and each of the source domain instances. Two example test instances (one positive and one negative) from the target domain **B**

T	uni-MS	Self	PL	Att
B	79.46	79.60	79.57	**79.68**
D	82.32	82.49	82.71	**82.96**
E	84.93	84.97	**85.30**	**85.30**
K	87.17	87.18	87.30	**87.48**

Table 4: Classification accuracies (%) across different steps of the proposed method, evaluated on **Chen2012**.

are shown in Figures 2 and 3. We observe that different source instances contribute to the predicted labels in different ways. As expected, in Figure 2a more positive source instances are selected using the relatedness map for a positive target instance, and Figure 3a more negative source instances are selected for a negative target instance. After training, we find that the proposed method identifies the level of importance of different source domains. Example (1) is closer to **D**, whereas Example (2) is closer to **E** with a very high value of θ. Figures 2c and 3c show that the instance specific contribution to the target instance. The proposed method also identifies the level of importance within the most relevant source domain. Figure 3 shows the actual reviews as the top-5 evidences from the source domains in Example (2). Negative labelled source training instance from **E**: *"Serious problem."* is the most important instance with the highest contribution of $\psi_i(x)\theta(x)$ to the decision.

4.4 Comparisons against Prior Work

Table 5 compares the proposed method against the following methods on **Blitzer2007** dataset.

SCL: Structural Correspondence Learning (Blitzer et al., 2006, 2007) is a single-source DA method, trained on the union of all source domains and tested on the target domain. We report the published results from Wu and Huang (2016).

SFA: Spectral Feature Alignment (Pan et al., 2010) is a single-source DA method, trained on the union of all source domains, and tested on the target domain. We report the published results from Wu and Huang (2016).

SST: Sensitive Sentiment Thesaurus (Bollegala et al., 2011; Bollegala et al., 2013) is the SoTA multi-source DA method on **Blitzer2007**. We report the published results from Bollegala et al. (2011).

SDAMS: Sentiment Domain Adaptation with Multiple Sources proposed by Wu and Huang (2016). We report the results from the original paper.

T	uni-MS	SCL	SFA	SST	SDAMS	AMN	Proposed
B	80.00	74.57	75.98	76.32	78.29	79.75	**83.50**
D	76.00	76.30	78.48	78.77	79.13	79.83	**80.50**
E	74.75	78.93	78.08	83.63*	**84.18****	80.92*	80.00*
K	85.25	82.07	82.10	85.18	**86.29**	85.00	86.00

Table 5: Classification accuracies (%) for the proposed method and prior work on **Blitzer2007**. Statistically significant improvements over **uni-MS** according to the Binomial exact test are shown by "*" and "**" respectively at $p = 0.01$ and $p = 0.001$ levels.

T	uni-MS	mSDA	DANN	MDAN	MoE	Proposed
B	79.46	76.98	76.50	78.63	79.42	**79.68**
D	82.32	78.61	77.32	80.65	**83.35**	82.96
E	84.93	81.98	83.81	85.34	**86.62**	85.30
K	86.71	84.26	84.33	86.26	**87.96**	87.48

Table 6: Classification accuracies (%) for the proposed method and prior work on **Chen2012**.

AMN: End-to-End Adversarial Memory Network (Li et al., 2017) is a single-source DA method, trained on the union of all source domains, and tested on the target domain. We report the published results from Ding et al. (2019).

In Table 6, we compare our proposed method against the following methods on **Chen2012**.

mSDA: Marginalized Stacked Denoising Autoencoders proposed by Chen et al. (2012). We report the published results from Guo et al. (2018).

DANN: Domain-Adversarial Neural Networks proposed by Ganin et al. (2016). We report the published results from Zhao et al. (2018).

MDAN: Multiple Source Domain Adaptation with Adversarial Learning proposed by Zhao et al. (2018). We report the published results from the original paper.

MoE: Mixture of Experts proposed by Guo et al. (2018). We report the published results from the original paper.

From Tables 5 and 6, we observe that the proposed method obtains the best classification accuracy on Books domain (**B**) in both settings, which is the domain with the smallest number of unlabelled instances. In particular, when the amount of training instances are small, pseudo-labelling and domain-attention in our proposed method play a vital role in multi-source UDA. Although **SDAMS** (in **Blitzer2007**) and **MoE** (in **Chen2012**) outperform the proposed method, the simplicity and the ability to provide explanations are attractive properties for a UDA method when applying in an industrial setting involving a massive number of source domains such as sentiment classification in E-commerce reviews.

5 Conclusions

We propose a multi-source UDA method that combines self-training with an attention module. In contrast to prior works that select pseudo-labelled instances based on prediction confidence of a predictor learnt from source domains, our proposed method uses similarity to the target domain during adaptation. Our proposed method reports competitive performance against previously proposed multi-source UDA methods on two splits on a standard benchmark dataset.

References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*, 1st edition. Chapman & Hall/CRC.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of 5th International Conference on Learning Representations (ICLR)*.

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)*, volume 1, pages 2–1. Citeseer.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2009. A theory of learning from different domains. *Machine Learning*, 79:151–175.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *NIPS 2006*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

D. Bollegala, D. Weir, and J. Carroll. 2013. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1719–1731.

Danushka Bollegala, David Weir, and John Carroll. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, pages 132–141, Portland, Oregon, USA. Association for Computational Linguistics.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57.

Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. 2012. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):18.

Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, pages 1627–1634, USA. Omnipress.

Xia Cui, Frans Coenen, and Danushka Bollegala. 2017. Tsp: Learning task-specific pivots for unsupervised domain adaptation. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, pages 754–771.

Xiao Ding, Qiankun Shi, Bibo Cai, Ting Liu, Yanyan Zhao, and Qiang Ye. 2019. Learning multi-domain adversarial neural networks for text classification. *IEEE Access*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th international conference on artificial intelligence and statistics (ICAI)*, pages 249–256.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML)*, pages 513–520, USA. Omnipress.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *Proc. AAAI Conference on Artificial Intelligence*.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4694–4703. Association for Computational Linguistics.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. 2004. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 180–191. VLDB Endowment.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Domain attention with an ensemble of experts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 643–653, Vancouver, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2237–2243. AAAI Press.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. Learning unsupervised word translations without adversaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, WWW '10, pages 751–760, New York, NY, USA. ACM.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Jeffery Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Ryan Rifkin and Aldebaro Klautau. 2004. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. 2005. To transfer or not to transfer. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS) workshop on Inductive Transfer: 10 Years Later*, volume 898, pages 1–4.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1044–1054. Association for Computational Linguistics.

Lex Razoux Schultz, Marco Loog, and Peyman Mohajerin Esfahani. 2018. Distance based source domain selection for sentiment classification. *arXiv preprint arXiv:1808.09271*.

Anders Søgaard. 2010. Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 205–208. Association for Computational Linguistics.

Sainbayar Sukhbaatar, arthur szlam arthur arthur arthur arthur, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc.

Fangzhao Wu and Yongfeng Huang. 2016. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 301–310.

Kun Zhang, Mingming Gong, and Bernhard Scholkopf. 2015. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 3150–3157. AAAI Press.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. 2018. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, pages 8568–8579.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541.