Exploiting Coarse-to-Fine Task Transfer for Aspect-Level Sentiment Classification*

Zheng Li,[1] Ying Wei,[2] Yu Zhang,[1] Xiang Zhang,[1] Xin Li[3]
[1]Hong Kong University of Science and Technology, Hong Kong
[2]Tencent AI Lab, Shenzhen, China
[3]The Chinese University of Hong Kong, Hong Kong
zlict@cse.ust.hk, judywei@tencent.com, yu.zhang.ust@gmail.com, xzhangax@ust.hk, lixin@se.cuhk.edu.hk

Abstract

Aspect-level sentiment classification (ASC) aims at identifying sentiment polarities towards aspects in a sentence, where the aspect can behave as a general *Aspect Category* (AC) or a specific *Aspect Term* (AT). However, due to the especially expensive and labor-intensive labeling, existing public corpora in AT-level are all relatively small. Meanwhile, most of the previous methods rely on complicated structures with given scarce data, which largely limits the efficacy of the neural models. In this paper, we exploit a new direction named *coarse-to-fine task transfer*, which aims to leverage knowledge learned from a rich-resource source domain of the coarse-grained AC task, which is more easily accessible, to improve the learning in a low-resource target domain of the fine-grained AT task. To resolve both the aspect granularity inconsistency and feature mismatch between domains, we propose a Multi-Granularity Alignment Network (MGAN). In MGAN, a novel Coarse2Fine attention guided by an auxiliary task can help the AC task modeling at the same fine-grained level with the AT task. To alleviate the feature false alignment, a contrastive feature alignment method is adopted to align aspect-specific feature representations semantically. In addition, a large-scale multi-domain dataset for the AC task is provided. Empirically, extensive experiments demonstrate the effectiveness of the MGAN.

Introduction

Aspect-level sentiment classification (ASC) aims to infer sentiment polarities over aspect categories (AC) or aspect terms (AT) distributed in sentences (Pang, Lee, and others 2008; Liu 2012). An aspect category implicitly appears in the sentence, which describes a general category of the entities. For example, in the sentence "*The salmon is tasty while the waiter is very rude*", the user speaks positively and negatively towards two aspect categories *"**food**"* and *"**service**"*, respectively. An aspect term characterizes a specific entity that explicitly occurs in the sentence. Considering the same sentence "*The **salmon** is tasty while the **waiter** is very rude*", the aspect terms are "***salmon***" and "***waiter***", and the user expresses positive and negative sentiments over them, respectively. In terms of the aspect granularity, the AC task is coarse-grained while the AT task is fine-grained.

To model aspect-oriented sentiment analysis, equipping Recurrent Neural Networks (RNNs) with the attention mechanism has become a mainstream approach (Tang et al. 2015; Wang et al. 2016; Ma et al. 2017; Chen et al. 2017), where RNNs aim to capture sequential patterns and the attention mechanism is to emphasize appropriate context features for encoding aspect-specific representations. Typically, attention-based RNN models can achieve good performance only when large corpora are available. However, AT-level datasets require the aspect terms to be comprehensively manually labeled or extracted by sequence labeling algorithms from the sentences, which is especially costly to obtain. Thus, existing public AT-level datasets are all relatively small, which limits the potential of neural models.

Nonetheless, we observe that plentiful AC-level corpora are more easily accessible. This is because that aspect categories are usually in a small set of general aspects that can be pre-defined. For example, commercial services such as review sites or social media can define a set of valuable aspect categories towards products or events in a particular domain (e.g., "*food*", "*service*", "*speed*", and "*price*" in the *Restaurant* domain). As a result, the mass collections of user preferences towards different aspect categories become practicable. Motivated by this observation, we propose a new problem named *coarse-to-fine task transfer* across both domain and granularity, with the aim of borrowing knowledge from an abundant source domain of the coarse-grained AC task to a small-scale target domain of the fine-grained AT task.

The challenges in fulfillment of this setting are two-fold: (1) task discrepancy: the two tasks concern with the aspects with different granularity. Source aspects are coarse-grained aspect categories, which lack a priori position information in the context. However, target aspects are fine-grained aspect terms, which have accurate position information. Thus, inconsistent granularity in aspects causes the discrepancy between tasks; (2) feature distribution discrepancy: generally the domains in the two tasks are different, which causes the distribution shift for both the aspects and its context between domains. For example, in the source *Restaurant* domain, *tasty* and *delicious* are used to express positive sentiment towards the aspect category "*food*", while *lightweight* and *responsive* often indicate positive sentiment towards the

*This work is supported by Hong Kong CERG grants (16209715 and 16244616), and NSFC (61473087 and 61673202).

aspect term "*mouse*" in the target *Laptop* domain.

To resolve the challenges, we propose a novel framework named **M**ulti-**G**ranularity **A**lignment **N**etwork (MGAN) to simultaneously align aspect granularity and aspect-specific feature representations across domains. Specifically, the MGAN consists of two networks for learning aspect-specific representations for the two domains, respectively. First, to reduce the task discrepancy between domains, i.e., modeling the two tasks at the same fine-grained level, we propose a novel Coarse2Fine (C2F) attention module to help the source task automatically capture the corresponding aspect term in the context towards the given aspect category (e.g., "salmon" to the "*food*"). Without any additional labeling, the C2F attention module can learn the coarse-to-fine process by an auxiliary task. Actually, more specific aspect terms and their position information are most directly pertinent to the expression of sentiment. The C2F module makes up these missing information for the source task, which effectively reduces the aspect granularity gap between tasks and facilitates the subsequent feature alignment.

Second, considering that a sentence may contain multiple aspects with different sentiments, thus capturing incorrect sentiment features towards the aspect can mislead feature alignment. To prevent false alignment, we adopt the Contrastive Feature Alignment (CFA) (Motiian et al. 2017) to semantically align aspect-specific representations. The CFA considers both semantic alignment by maximally ensuring the equivalent distributions from different domains but the same class, and semantic separation by guaranteeing distributions from both different classes and domains to be as dissimilar as possible. Moreover, we build a large-scale multi-domain dataset named *YelpAspect* with 100K samples for each domain to serve as highly beneficial source domains. Empirically, extensive experiments demonstrate that the proposed MGAN model can achieve superior performances on two AT-level datasets from SemEval'14 ABSA challenge and an ungrammatical AT-level twitter dataset.

Our contributions of this paper are four-fold: (1) to the best of our knowledge, a novel transfer setting cross both domain and granularity is first proposed for aspect-level sentiment analysis; (2) a new large-scale, multi-domain AC-level dataset is constructed; (3) the novel Coarse2Fine attention is proposed to effectively reduce the aspect granularity gap between tasks; (4) empirical studies verify the effectiveness of the proposed model on three AT-level benchmarks.

Related Work

Traditional supervised learning algorithms highly depend on extensive handcrafted features to solve aspect-level sentiment classification (Jiang et al. 2011; Kiritchenko et al. 2014). These models fail to capture semantic relatedness between the aspect and its context. To overcome this issue, the attention mechanism, which has been successfully applied in many NLP tasks (Bahdanau, Cho, and Bengio 2014; Sukhbaatar et al. 2015; Yang et al. 2016; Shen et al. 2017), can help the model explicitly capture intrinsic aspect-context association (Tang et al. 2015; Tang, Qin, and Liu 2016; Wang et al. 2016; Ma et al. 2017; Chen et al. 2017; Ma, Peng, and Cambria 2018; Li et al. 2018a). However, most of these methods highly rely on data-driven RNNs or tailor-made structures to deal with complicated cases, which requires substantial AT-level data to train effective neural models. Different from them, the proposed model can highly benefit from useful knowledge learned from a related abundant domain of the AC-level task.

Existing domain adaptation tasks for sentiment analysis focus on traditional sentiment classification without considering the aspect (Blitzer, Dredze, and Pereira 2007; Pan et al. 2010; Glorot, Bordes, and Bengio 2011; Chen et al. 2012; Bollegala, Weir, and Carroll 2013; Yu and Jiang 2016; Li et al. 2017; 2018b). In terms of data scarcity and the value of task, transfer learning is more urgent for aspect-level sentiment analysis that characterizes users' different preferences. To the best of our knowledge, only a few studies have explored to transfer from a single aspect category to another in a same domain based on adversarial training (Zhang, Barzilay, and Jaakkola 2017). Different from that, we explore a motivated and challenging setting which aims to transfer cross both aspect granularity and domain.

Multi-Granularity Alignment Network

In this section, we introduce the proposed MGAN model. We first present the problem definition and notations, followed by an overview of the model. Then we detail the model with each components.

Problem Definition and Notations

Coarse-to-fine task transfer Suppose that we have sufficient AC-level labeled data $\mathbf{X}^s = \{(\mathbf{x}_k^s, \mathbf{a}_k^s), y_k^s\}_{k=1}^{N^s}$ in a source domain D_s, where y_k^s is the sentiment label for the k-th sentence-aspect pair $(\mathbf{x}_k^s, \mathbf{a}_k^s)$. Besides, only a small amount of AT-level labeled data $\mathbf{X}^t = \{(\mathbf{x}_{k'}^t, \mathbf{a}_{k'}^t), y_{k'}^t\}_{k'=1}^{N^t}$ is available in a target domain D_t. Note that each source aspect $\mathbf{a}_k^s$ belongs to a set of pre-defined aspect categories C while each target aspect $\mathbf{a}_{k'}^t$ is a sub-sequence of $\mathbf{x}_{k'}^t$, i.e., aspect term. The goal of this task is to learn an accurate classifier to predict the sentiment polarity of target testing data.

An Overview of the MGAN model

The goal of the MGAN aims to transfer from a rich-resource source domain of an AC task to facilitate a low-resource target domain of an AT task. The architecture of the proposed MGAN is shown in Figure 1. Specifically, the MGAN consists of two networks for tackling the two aspect-level tasks respectively. To reduce the task discrepancy, the two networks contain different numbers of attention hops such that they can keep a consistent granularity and the symmetric information towards the aspect. In MGAN, two basic hop units are used similarly as common attention-based RNN models, where the Context2Aspect (**C2A**) attention aims to measure the importance of each aspect word and generate the aspect representation with the aid of each context word, and the Position-aware Sentiment (**PaS**) attention utilizes the obtained aspect representation and the position information of the aspect to capture relevant sentiment features in the context for encoding the aspect-specific representation.

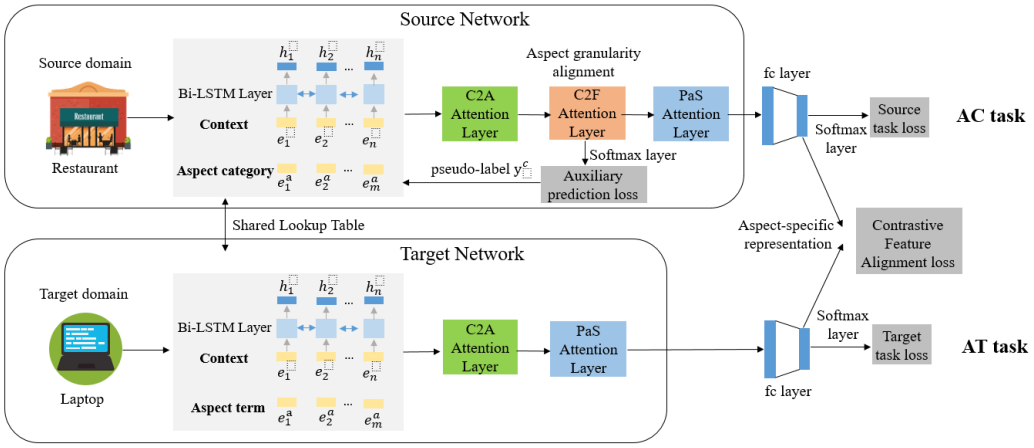

Figure 1: The architecture of the Multi-Granularity Alignment Network (MGAN) model.

Moreover, we build a Coarse2Fine (**C2F**) attention upon the C2A module to specifically model the source aspect before feeding to the PaS module. The C2F module uses the source aspect representation to attend corresponding aspect terms in the context and then the attended context features is conversely predicted the category of the source aspect (pseudo-label). After obtaining aspect-specific representations, the knowledge transfer between the two tasks is via the contrastive feature alignment. In summary, the source network acts as a "teacher", which consists of three-level attention hops (C2A+C2F+PaS) for the AC task, while the target network is like a "student" that only uses two basic attention hops (C2A+PaS) for the AT task. In the following sections, we introduce each component of MGAN in details.

Bi-directional LSTM layer

Given a sentence-aspect pair ($\mathbf{x}$, $\mathbf{a}$) from the source or target domain, we assume that the sentence consists of n words, i.e., $\mathbf{x}=\{w_1, w_2, ..., w_n\}$, and the aspect contains m words, i.e., $\mathbf{a}=\{w_1^a, w_2^a, ..., w_m^a\}$. Then we map them into its embedding vectors $\mathbf{e}=\{\mathbf{e}_i\}_{i=1}^n \in \mathbb{R}^{n\times d_w}$ and $\mathbf{e}^a=\{\mathbf{e}_j^a\}_{j=1}^m \in \mathbb{R}^{m\times d_w}$ respectively. To capture phrase-level sentiment features in the context (e.g., "not satisfactory"), we employ a Bi-directional LSTM (Bi-LSTM) to preserve the contextual information for each word of the input sentence. The Bi-LSTM transforms the input $\mathbf{e}$ into the contextualized word representations $\mathbf{h}=\{\mathbf{h}_i\}_{i=1}^n \in \mathbb{R}^{n\times 2d_h}$ (i.e. hidden states of Bi-LSTM). For simplicity, we denote the operation of an LSTM unit on $\mathbf{e}_i$ as LSTM($\mathbf{e}_i$). Thus, the contextualized word representation $\mathbf{h}_i \in \mathbb{R}^{2d_h}$ is obtained as

$$\mathbf{h}_i = [\overrightarrow{\text{LSTM}}(\mathbf{e}_i); \overleftarrow{\text{LSTM}}(\mathbf{e}_i)], i \in [1, n], \quad (1)$$

where ";" denotes the vector concatenation.

Context2Aspect (C2A) Attention

Not all aspect words contribute equally to the semantic of the aspect. For example, in the aspect term "*techs at HP*", the sentiment is usually expressed over the headword "*techs*" but seldom over modifiers like the brand name "*HP*". Thus, "*techs*" is more important than "*at*" and "*HP*". This also applies to the aspect category (e.g., "*food seafood fish*"). Thus, we propose the Context2Aspect (C2A) attention to measure the importance of the aspect words with regards to each context word. Formally, we calculate a pair-wise alignment matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ between the context and the aspect, where the alignment score $M(i, j)$ between the i-th context word and the j-th aspect word is obtained as

$$M(i, j) = \tanh(\mathbf{W}_a[\mathbf{h}_i; \mathbf{e}_j^a] + b_a), \quad (2)$$

where $\mathbf{W}_a$ and b_a are learnable parameters. Then, we apply a row-wise softmax function to get probability distributions in each row. By defining $\boldsymbol{\delta}(i) \in \mathbb{R}^m$ as the individual aspect-level attention given the i-th context word, we average all the $\boldsymbol{\delta}(i)$'s to get the C2A attention as $\boldsymbol{\alpha} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}(i)$. The C2A attention further contributes the context-aware aspect representation by $\mathbf{h}_*^a = \sum_{j=1}^m \alpha_j \mathbf{e}_j^a$, where $* \in \{s, t\}$ denotes the source or target domain. We tackle the aspect representation $\mathbf{h}_*^a$ for the two tasks differently, where $\mathbf{h}_s^a$ is fed to the C2F module while $\mathbf{h}_t^a$ is directly fed to the PaS module.

Coarse2Fine (C2F) Attention

Aspect terms, which act as the true "opinion entity", are the most directly pertinent to the expression of sentiment. However, source task concerns with coarse-grained aspect categories that lack of detailed position information in the context. We wish to achieve task alignment such that the target task can leverage more useful knowledge learned from the source task at the same fine-grained level. It is observed that the number of source aspects is much smaller and many instances contain same aspect category, but the underlying entities can behave diversely in different contexts. For example, the aspect category *"food seafood fish"* can be instantiated as "salmon", "tuna", "taste" and etc.

Based on this observation, we can capture more specific semantics of the source aspect and its position information conditioned on its context. Motivated by autoencoders (Bengio et al. 2007), we introduce an auxiliary pseudo-label prediction task for the source task. In this task, a source aspect $\mathbf{a}^s$ is not only regarded as a sequence of aspect words,

but also as a pseudo-label (category of the aspect) y^c, where $c \in C$ and C is a set of aspect categories. We utilize the obtained aspect representation $\mathbf{h}_s^a$ for $\mathbf{a}^s$ to attend the context and then the induced attention scores aggregate the context information to conversely predict the pseudo category label of $\mathbf{a}^s$ itself. If the context contains the aspect term correlated closely to the source aspect, then the attention mechanism can emphasize it for better prediction. We denote this mechanism as Coarse2Fine attention, which is calculated as:

$$z_i^f = (\mathbf{u}_f)^T \tanh(\mathbf{W}_f[\mathbf{h}_i; \mathbf{h}_s^a] + \mathbf{b}_f), \tag{3}$$

$$\beta_i^f = \frac{\exp(z_i^f)}{\sum_{i'=1}^{n} \exp(z_{i'}^f)}, \tag{4}$$

$$\mathbf{v}^a = \sum_{i=1}^{n} \beta_i^f \mathbf{h}_i, \tag{5}$$

where $\mathbf{W}_f \in \mathbb{R}^{d_u \times (2d_h + d_e)}$, $\mathbf{b}_f \in \mathbb{R}^{d_u}$ and $\mathbf{u}_f \in \mathbb{R}^{d_u}$ are the weights of the layer. We feed the attended representation $\mathbf{v}^a$ to a *softmax* layer for the auxiliary task prediction, which is trained by minimizing the cross-entropy loss between the predicted pseudo-label $\hat{y}_k^c$ and its ground-truth y_k^c as:

$$\mathcal{L}_{aux} = -\frac{1}{N_s} \sum_{k=1}^{N_s} \sum_{c \in C} y_k^c \log \hat{y}_k^c. \tag{6}$$

However, there may not exist corresponding aspect term when the context implicitly expresses a sentiment toward the aspect category. To overcome this issue, similar to the gate mechanism in RNN variants (Jozefowicz, Zaremba, and Sutskever 2015), we adopt a fusion gate $\mathbf{F}$ to adaptively controls the passed proportions of $\mathbf{h}_s^a$ and $\mathbf{v}^a$ towards a more specific source aspect representation $\mathbf{r}_s^a$:

$$\mathbf{F} = \text{sigmoid}(\mathbf{W}[\mathbf{v}^a; \mathbf{h}_s^a] + \mathbf{b}), \tag{7}$$

$$\mathbf{r}_s^a = \mathbf{F} \odot \mathbf{h}_s^a + (\mathbf{1} - \mathbf{F}) \odot \mathbf{W}' \mathbf{v}^a, \tag{8}$$

where $\mathbf{W} \in \mathbb{R}^{d_e \times (d_e + 2d_h)}$ and $\mathbf{b} \in \mathbb{R}^{d_e}$ are the weights of the gate, $\mathbf{W}' \in \mathbb{R}^{d_e \times 2d_h}$ performs dimension reduction, and $\odot$ denotes element-wise multiplication.

Position-aware Sentiment (PaS) Attention

Following an important observation found in (Tang, Qin, and Liu 2016; Chen et al. 2017) that a closer sentiment word is more likely to be the actual modifier of the aspect term (e.g., in "*great food but the service is dreadful*", "*great*" is more closer to "*food*" than "*service*"), we take the position information of the aspect term into consideration for designing the PaS attention. For the target domain, we adopt a proximity strategy to calculate the target position relevance between the i-th context word and aspect term as follows:

$$p_i^t = \begin{cases} 1 - \frac{m_0 - i}{n} & i < m_0 \\ 0 & m_0 \le i \le m_0 + m \\ 1 - \frac{i - (m_0 + m)}{n} & i > m_0 + m \end{cases}, \tag{9}$$

where m_0 is the index of the first aspect word, n and m are the length of the sentence and aspect, respectively.

Unfortunately, in the source domain where aspect category is given, the exact position of the corresponding aspect term is not directly accessible. Instead, the C2F attention vector $\boldsymbol{\beta}^f \in \mathbb{R}^n$, indicating the probability of each context word being an aspect term, can help establish the position relevance. We first define a location matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$ to represent the proximity of each word in the sentence:

$$L_{ii'} = 1 - \frac{|i - i'|}{n}, i, i' \in [1, n]. \tag{10}$$

Then we calculate the source position relevance for i-th context word with the aid of C2F attention weights by $p_i^s = \mathbf{L}_i \boldsymbol{\beta}^f$. Obviously, the i-th context word closer to a possible aspect term with a large value in $\boldsymbol{\beta}^f$ will have a larger position relevance p_i^s. Finally, the PaS attention is calculated by a general form for both domains:

$$z_i^o = (\mathbf{u}_o)^T \tanh(\mathbf{W}_o[\mathbf{h}_i; \mathbf{r}_*^a] + \mathbf{b}_o), \tag{11}$$

$$\gamma_i^o = \frac{\exp(p_i^* z_i^o)}{\sum_{i'=1}^{n} \exp(p_{i'}^* z_{i'}^o)}, \tag{12}$$

$$\mathbf{v}^o = \sum_{i=1}^{n} \gamma_i^o \mathbf{h}_i, \tag{13}$$

where p_i^* is the position relevance and $\mathbf{r}_*^a$ is the input aspect representation, with $* \in \{s, t\}$ denoting the source or target domain (note that $\mathbf{r}_t^a = \mathbf{h}_t^a$). Then we pass the aspect-specific representation $\mathbf{v}^o$ to a fully-connected layer and a *softmax* layer for sentiment classification. The sentiment classification tasks for both domain are trained by minimizing two cross-entropy losses $\mathcal{L}_{sen}^s$ and $\mathcal{L}_{sen}^t$, respectively.

Contrastive Feature Alignment

After obtaining aspect-specific representations of two domains at the same granularity, we would further bridge the distribution gap across domains. The prevalent unsupervised domain adaptation methods (Gretton et al. 2007; Ganin et al. 2016) require enormous unlabeled target data to achieve satisfactory performances, which is impractical in our problem where collecting unlabeled data needs labor-intensive annotations of all aspect terms in the sentences. Therefore, inspired by (Motiian et al. 2017), we perform Contrastive Feature Alignment (CFA) by fully utilizing the limited target labeled data to semantically align representations across domains. Mathematically, we parameterize the two networks by g_s and g_t, and denote the probability distribution by $\mathbb{P}$. Specifically, the CFA consisits of semantic alignment (SA) and semantic separation (SS). The SA aims to ensure identical distributions of feature representations $\mathbb{P}(g_s(\mathbf{X}^s))$ and $\mathbb{P}(g_t(\mathbf{X}^t))$ conditioned on different domains but the same class, while the SS further alleviates false alignment by guaranteeing $\mathbb{P}(g_s(\mathbf{X}^s))$ and $\mathbb{P}(g_t(\mathbf{X}^t))$ to be as dissimilar as possible conditioned on both different domains and classes. Considering that only a small amount of target labeled data is available, we revert the CFA characterizing distributions with enough data to pair-wise surrogates as:

$$\mathcal{L}_{cfa} = \sum_{k,k'} \omega(g_s(\mathbf{x}_k^s, \mathbf{a}_k^s), g_t(\mathbf{x}_{k'}^t, \mathbf{a}_{k'}^t)), \tag{14}$$

where $\omega(\cdot, \cdot)$ is a contrastive function that performs semantic alignment or separation in terms of supervised information from both domains. Formally, $\omega(\cdot, \cdot)$ is defined as:

$$\omega(\mathbf{u}, \mathbf{v}) = \begin{cases} \|\mathbf{u} - \mathbf{v}\|^2 & \text{if} \quad y_k^s = y_{k'}^t, \\ \max(0, D - \|\mathbf{u} - \mathbf{v}\|^2) & \text{if} \quad y_k^s \neq y_{k'}^t, \end{cases} \tag{15}$$

where D is a parameter dictating the degree of separation and is set to 1 in our experiments.

Alternating Training

Combining the losses we introduced before together with a ℓ_2 regularization, we constitute the overall losses for the source and target networks as:

$$\mathcal{L}_{src} = \mathcal{L}^s_{sen} + \mathcal{L}_{aux} + \lambda\mathcal{L}_{cfa} + \rho\mathcal{L}^s_{reg}, \quad (16)$$

$$\mathcal{L}_{tar} = \mathcal{L}^t_{sen} + \lambda\mathcal{L}_{cfa} + \rho\mathcal{L}^t_{reg}, \quad (17)$$

where λ, ρ balance the effect of the CFA loss and the ℓ_2 regularization loss, respectively. The source network has one more auxiliary loss $\mathcal{L}_{aux}$ compared with the target one to achieve task alignment. The whole training procedure consists of two stages: (1) To prevent early overfitting of the target domain, the source network S is individually trained on the source domain by optimizing $\mathcal{L}^s_{sen} + \mathcal{L}_{aux} + \rho\mathcal{L}^s_{reg}$. Then, S and the BiLSTM, C2A, and PaS modules of S are used to initialize the source and target networks of the MGAN, respectively. (2) We alternately optimize $\mathcal{L}_{src}$ for the source network and $\mathcal{L}_{tar}$ for the target network.

Experiments

Datasets

Source: AC-level We build a large-scale, multi-domain dataset named *YelpAspect* as source domains, which is obtained similarly as the Yelp recommendation dataset (Bauman, Liu, and Tuzhilin 2017). Specifically, YelpAspect contains three domains: Restaurant (**R1**), Beautyspa (**B**), and Hotel (**H**). The statistics of the YelpAspect dataset are summarized in Table 1. Yelp reviews are collected in US cities over six years. Aspect categories and sentiment labels are identified by the "industrial-strength" Opinion Parser (OP) system (Qiu et al. 2011; Liu 2015). To be consistent with the target domain datasets, YelpAspect is preprocessed in the sentence level by OP, while the dataset in (Bauman, Liu, and Tuzhilin 2017) is in the document level. Moreover, we manually double-check to correct wrong annotations produced by OP system and purposely select more negation, contrastive and question instances to make it more challenging. The dataset is available at https://github.com/hsqmlzno1/MGAN.

Source domain		#Pos	#Neu	#Neg	#Asp
Restaurant (R1)	Train	46,315	45,815	16,020	68
	Test	5,207	4,944	1,743	
Beautyspa (B)	Train	45,770	42,580	16,023	45
	Test	5,056	4,793	1,823	
Hotel (H)	Train	40,775	36,901	20,864	44
	Test	4,418	4,048	2,450	

Table 1: The YelpAspect dataset. #Asp denotes the number of aspect categories.

Target: AT-level For target domains, we use three public benchmark datasets: Laptop (**L**), Restaurant (**R2**) and Twitter (**T**). The Laptop and Restaurant are from SemEval'14 ABSA challenge (Kiritchenko et al. 2014) by removing a few examples which have "conflict labels" as done in (Chen et al. 2017). The Twitter dataset is collected by (Dong et al. 2014), containing ungrammatical twitter posts. Table 2 summarizes the statistics of the target domain datasets.

Target Domain		#Pos	#Neu	#Neg
Laptop (L)	Train	980	454	858
	Test	340	171	128
Restaurant (R2)	Train	2,159	632	800
	Test	730	196	195
Tweets (T)	Train	1,567	3,127	1,563
	Test	174	346	174

Table 2: Statistics of the target domain datasets.

Experimental Setup

To evaluate our proposed method, we construct eight coarse-to-fine transfer tasks: R1→L, H→L, B→L, H→R2, B→R2, R1→T, H→T, B→T, where we do not use the pair (R1, R2) as they come from the same domain. For each transfer pair $D_s \to D_t$, the training data from domain D_s and randomly sampled 90% training data from domain D_t are used for training, the rest 10% training data from D_t is used for validation, and the testing data from D_t is used for testing. Evaluation metrics are Accuracy and Macro-Average F1, where the latter is more suitable for imbalanced datasets.

Implementation Details

The word embeddings are initialized with 200-dimension GloVE vectors (Pennington, Socher, and Manning 2014) and fine-tuned during the training. d_e, d_h, d_u are set to be 200, 150 and 100, respectively. The fc layer size is 300. The Adam (Kingma and Ba 2014) is used as the optimizer with the initial learning rate 10^{-4}. Gradients with the ℓ_2 norm larger than 40 are normalized to be 40. All weights in networks are randomly initialized from a uniform distribution $U(-0.01, 0.01)$. The batch sizes are 64 and 32 for source and target domains, respectively. The control-off factors λ, ρ are set to be 0.1 and 10^{-6}. To alleviate overfitting, we apply dropout on the word embeddings of the context with dropout rate 0.5. We also perform early stopping on the validation set during the training process. The hyperparameters are tuned on 10% randomly held-out training data of the target domain in R1→L task and are fixed to be used in all transfer pairs.

Baseline Methods

The baseline methods are divided into two groups:

Non-Transfer To demonstrate the benefits from coarse-to-fine task transfer, we compare with the following state-of-the-art AT-level methods without transfer:

- **TD-LSTM** (Tang et al. 2015): It employs two LSTMs to model the left and right contexts of the aspect and then concatenates the context representations for prediction.
- **AE-LSTM**, and **ATAE-LSTM** (Wang et al. 2016): AE-LSTM is a simple LSTM model incorporating the aspect embedding as input, while ATAE-LSTM extends AE-LSTM with the attention mechanism.
- **MemNet** (Tang, Qin, and Liu 2016): it applies a memory network with multi-hops attentions and predicts sentiment based on the top-most context representations.
- **IAN** (Ma et al. 2017): It adopts two LSTMs to learn the representations of the context and the aspect interactively;

	Model	L		R2		T	
		Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Baselines	AE-LSTM (Wang et al. 2016)	68.97	62.50	76.25	64.32	69.42	56.28
	ATAE-LSTM (Wang et al. 2016)	68.65	62.45	77.23	64.95	69.58	56.72
	TD-LSTM (Tang et al. 2015)	68.18	62.28	75.63	64.16	66.62	64.01
	IAN (Ma et al. 2017)	72.10	-	78.60	-	-	-
	MemNet (Tang, Qin, and Liu 2016)	70.33	64.09	78.16	65.83	68.50	66.91
	RAM (Chen et al. 2017)	72.08	68.43	78.48	68.54	69.36	67.30
Base Model	TN	70.58	65.34	77.91	65.75	71.68	71.02
		Average results over each target domain					
Ablated Models	MGAN w/o PI	72.98	67.71	78.99	66.41	72.88	71.57
	MGAN w/o C2F	74.80	69.63	80.46	67.86	73.53	72.37
Full Model	MGAN	**76.21**†‡	**71.42**†‡	**81.49**†‡	**71.48**†‡	**74.62**†‡	**73.53**†‡

Table 3: Experimental results (%). The marker † refers to p-value < 0.05 when comparing with MGAN w/o C2F , while the marker ‡ refers to p-value < 0.05 when comparing with RAM.

- **RAM** (Chen et al. 2017): It employs multiple attentions with a GRU cell to non-linearly combine the aggregation of word features in each layer.
- **Target Network (TN)**: It is our proposed base model (BiLSTM+C2A+Pas) trained on D_t for the target task.

For IAN, we report the results in the original paper and use the source codes of other methods for experiments.

Transfer To investigate the effectiveness of the CFA , we also compare the following transfer methods:

- **Source-only (SO)**: It uses a source network trained on D_s to initialize a target network and then tests it on D_t.
- **Fine-tuning (FT)**: It advances SO with further fine-tuning the target network on D_t.
- **M-DAN**: It is a multi-adversarial version of Domain Adversarial Network (DAN) (Ganin et al. 2016) based on multiple domain discriminators. All discriminators are built upon the PaS layers of the two networks, each of which aligns one class distribution between domains.
- **M-MMD**: Similar with M-DAN, M-MMD aligns different class distributions between domains based on multiple Maximum Mean Discrepancy (MMD) (Gretton et al. 2007). For each MMD, following the (Bousmalis et al. 2016), we use a linear combination of 19 RBF kernels with the width parameters ranging from 10^{-6} to 10^{6}.

The original DAN and MMD are unsupervised domain adaptation methods. Thus, for fair comparison, we use the source code of DAN and MMD, and extend them to M-DAN and M-MMD that utilize target supervised information and have higher performances, respectively.

Result Analysis

Comparison with Non-Transfer Note that we are the first to explore transfer techniques and achieve the best performances in this task. Thus, it is necessary to show our improvements over current superior non-transfer methods. The classification results are shown in Table 3. The results of our full model and its ablations are calculated by averaging over each target domain among eight transfer pairs (e.g., R2 is obtained by averaging over H→R2 and B→R2). Based on the table, we have the following observations: (1) Our full model MGAN consistently and significantly achieves the best results in all target domains, outperforming the strongest baseline RAM by 4.13%, 3.58%, 5.26% for accuracy and 2.99%, 2.94% and 6.23% for Macro-F1 on average. Our base model TN that does not utilize the knowledge from the source task, can only compete against with the baselines. It could be more convincing that the MGAN can achieve superior performances even with a simple model for the target task. This also indicates that the efficacy of MGAN benefits from leveraging useful knowledge learned from the source task. (2) MGAN consistently outperforms the MGAN w/o C2F, where C2F module of the source network is removed and the source position information is missed (we set all p_i^s to 1), by 1.41%, 1.03%, 1.09% for accuracy and 1.79%, 3.62% and 1.16% for Macro-F1 on average. This is because that the C2F can effectively reduce the aspect granularity gap between tasks such that more useful knowledge can be distilled to facilitate the target task. (3) Position information is crucial for aspect-level sentiment analysis. The MGAN w/o PI, which does not utilize the position information, performs very poorly.

Acc / Macro-F1	SO	FT	M-DAN	M-MMD	MGAN w/o SS	MGAN
R1→L	69.80	74.80	75.74	75.90	77.00	**77.62**
	67.05	69.84	71.13	70.95	71.31	**72.26**
B→L	70.27	71.99	72.46	74.02	74.49	**75.74**
	66.84	67.13	68.69	69.36	69.94	**71.65**
H→L	70.74	72.77	75.43	73.71	74.02	**75.27**
	67.89	67.75	71.40	69.16	69.31	**70.34**
B→R2	72.90	79.16	79.96	81.31	**81.84**	81.66
	64.36	66.78	68.73	70.54	**71.80**	71.72
H→R2	72.36	80.59	79.87	79.87	80.95	**81.31**
	62.48	69.57	69.19	67.58	70.57	**71.24**
R1→T	46.39	72.83	72.11	73.41	73.41	**75.00**
	45.74	72.10	70.69	72.52	72.76	**74.00**
B→T	46.39	72.25	72.98	73.27	73.27	**74.00**
	45.62	70.30	71.88	72.34	71.79	**72.87**
H→T	47.40	71.82	72.55	73.27	73.99	**74.86**
	46.71	70.05	71.07	72.11	72.32	**73.73**
Average	62.03	74.53	75.13	75.60	76.12	**76.93**†
	58.34	69.19	70.33	70. 57	71.23	**72.23**†

Table 4: Experimental results (%). The marker † refers to p-value < 0.05 when comparing with MGAN w/o SS.

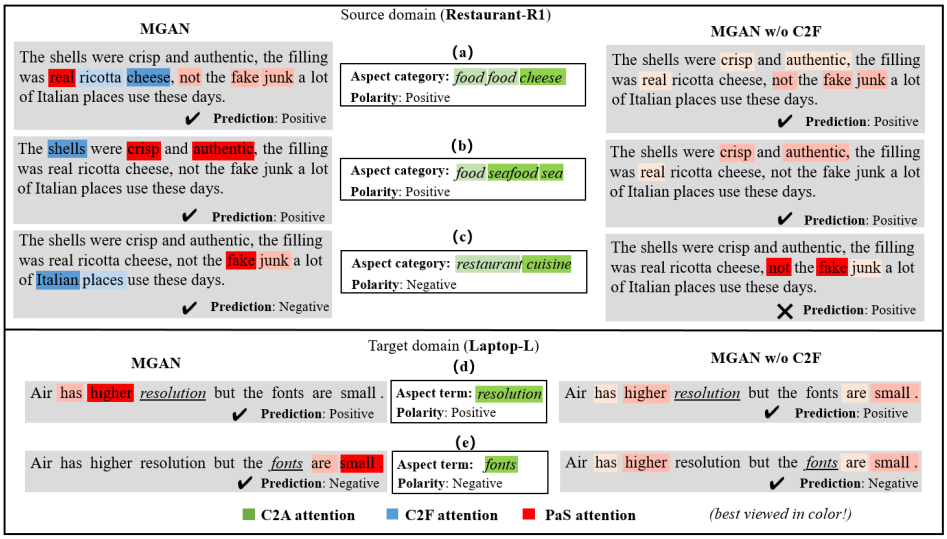

Figure 2: Visualization of attention: MGAN versus MGAN w/o C2F in the R1→L task. Deeper color denotes higher weights.

Aspect category: ***food dessert***

cake crepe cookie syrup sorbet sundae cannoli custard cupcakes chocolate vheesecake ice cream biscuit brownie pumpkin pie pudding muffin macaroon

Aspect category: ***restaurant interior***

space curtains wall seat table chair decor flooring window area lounge board decorations furniture desk screen ceiling door wallpaper patio bathroom lighting floor restroom

Aspect category: ***service***

hostess bartender server serving owner cashier staff employee service waitress waitstaff waiter manager management inspector

Aspect category: ***food seafood sea***

crab clam calamari sea shell shrimp scallop lobster prawns mussel linguine oyster shells clams

Figure 3: Associated aspect terms towards different aspect categories captured by C2F attention in the R1→L task.

Comparison with Transfer To avoid the effect of aspect granularity gap, all these models keep the C2F module. The compared results are shown in Table 4. SO performs poorly due to no adaptation applied. The popular technique FT cannot achieve satisfactory results since fine-tuning may cause the oblivion of useful knowledge from the source task. The full model MGAN outperforms M-DAN and M-MMD by 1.80% and 1.33% for accuracy and 1.90% and 1.66% for Marco-F1 on average, respectively. We derive two possible reasons: First, enormous target data is unavailable since it is hard to obtain, thus, it may be insufficient to represent target distributions by limited target labeled data for the distribution alignment based methods; Second, M-DAN and M-MMD focus on the semantic alignment but ignore semantic separation. Remarkably, MGAN considers both of them in a point-wise surrogate, which altogether improves the performance of our method. Besides, MGAN outperforms its ablation MGAN w/o SS removing the semantic separation loss of the CFA by 0.81% for accuracy and 1.00% for Macro-F1 on average, which implies that the semantic separation plays an important role in alleviating false alignment.

Effect of C2F Attention Module

We now give some illustrated examples to show the effect of C2F for solving aspect granularity inconsistency, by comparing MGAN and MGAN w/o C2F. Some hard cases containing multiple sentiment-aspect paris in the R1→L task are shown in Figure 2. In the source domain R1, both models first utilize the C2A to attend the informative part of the aspect category, e.g., *"cheese"*, *"seafood sea"* and *"cuisine"*, which are representatives for each aspect. Then, compared with MGAN w/o C2F, MGAN further uses C2F to capture more specific aspect terms from the context towards the aspect category, such as "shells" to *food seafood sea*, which helps the source task capture more fine-grained semantics of aspect category and detailed position information like the target task, such that the sentiment attention can be position-aware and identify more relevant sentiment features towards the aspect. For example, in the (a) and (c), the user expresses a positive sentiment over *food food cheese* but a negative attitude towards *restaurant cuisine* (*cuisine* means a style of cooking especially as a characteristic of a particular country or region). MGAN captures the regional words for the cooking style, i.e., "italian place" towards *restaurant cuisine* and the related n-gram sentiment feature "fake junk" instead of the "not the fake junk" for the "ricotta cheese", and finally makes a correct prediction, which helps distill more useful knowledge for subsequent feature alignment. While MGAN w/o C2F locates wrong sentiment contexts and fails in (c). As such, benefited from distilled knowledge from the source task, MGAN can better model the complicated relatedness between the context and aspect term for the target domain L, but MGAN w/o C2F performs poorly though it make true predictions in (d) and (e). Moreover, as shown in Figure 3, we list some samples of captured associated aspect terms towards different aspect categories based on the highest C2F attention weight. These underlying aspect terms make the source task more correlated to the target task.

Conclusion and Future Work

In this paper, we explore a motivated direction for aspect-level sentiment classification named *coarse-to-fine task transfer* and build a large-scale YelpAspect dataset as highly beneficial source benchmarks. A novel MGAN model is proposed to solve both aspect granularity inconsistency and domain feature mismatch problems, and achieves superior performances. Moreover, there are many other potential directions, like transferring between different aspect categories across domains, transferring to a AT-level task where the aspect terms are also not given and need to be firstly identified. We believe all these can help improve the ASC task and there will be more effective solutions coming in the near future.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bauman, K.; Liu, B.; and Tuzhilin, A. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *KDD*, 717–725. ACM.

Bengio, Y.; Lamblin, P.; Popovici, D.; and Larochelle, H. 2007. Greedy layer-wise training of deep networks. In *NIPS*, 153–160.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.

Bollegala, D.; Weir, D.; and Carroll, J. 2013. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *TKDE* 25(8):1719–1731.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *NIPS*, 343–351.

Chen, M.; Xu, Z.; Sha, F.; and Weinberger, K. Q. 2012. Marginalized denoising autoencoders for domain adaptation. In *ICML*, 767–774.

Chen, P.; Sun, Z.; Bing, L.; and Yang, W. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *EMNLP*, 452–461.

Dong, L.; Wei, F.; Tan, C.; Tang, D.; Zhou, M.; and Xu, K. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *ACL*, volume 2, 49–54.

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

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 513–520.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A kernel method for the two-sample-problem. In *NIPS*, 513–520.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. 2011. Target-dependent twitter sentiment classification. In *ACL-HLT*, 151–160. Association for Computational Linguistics.

Jozefowicz, R.; Zaremba, W.; and Sutskever, I. 2015. An empirical exploration of recurrent network architectures. In *ICML*, 2342–2350.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiritchenko, S.; Zhu, X.; Cherry, C.; and Mohammad, S. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *SemEval*, 437–442.

Li, Z.; Zhang, Y.; Wei, Y.; Wu, Y.; and Yang, Q. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, 2237.

Li, X.; Bing, L.; Lam, W.; and Shi, B. 2018a. Transformation networks for target-oriented sentiment classification. In *ACL*, 946–956.

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

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

Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Ma, D.; Li, S.; Zhang, X.; and Wang, H. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Ma, Y.; Peng, H.; and Cambria, E. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *AAAI*.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified deep supervised domain adaptation and generalization. In *ICCV*, volume 2, 3.

Pan, S. J.; Ni, X.; Sun, J.-T.; Yang, Q.; and Chen, Z. 2010. Cross-domain sentiment classification via spectral feature alignment. In *WWW*, 751–760. ACM.

Pang, B.; Lee, L.; et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Qiu, G.; Liu, B.; Bu, J.; and Chen, C. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics* 37(1):9–27.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2017. Disan: Directional self-attention network for rnn/cnn-free language understanding. *arXiv preprint arXiv:1709.04696*.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *NIPS*, 2440–2448.

Tang, D.; Qin, B.; Feng, X.; and Liu, T. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.

Wang, Y.; Huang, M.; Zhao, L.; et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *EMNLP*, 606–615.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *NAACL-HLT*, 1480–1489.

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

Zhang, Y.; Barzilay, R.; and Jaakkola, T. 2017. Aspect-augmented adversarial networks for domain adaptation. *arXiv preprint arXiv:1701.00188*.