Coarse Alignment of Topic and Sentiment: A Unified Model for Cross-Lingual Sentiment Classification

Deqing Wang, Baoyu Jing, Chenwei Lu, Junjie Wu, Guannan Liu, Chenguang Du, and Fuzhen Zhuang

***Abstract*— Cross-lingual sentiment classification (CLSC) aims to leverage rich-labeled resources in the source language to improve prediction models of a resource-scarce domain in the target language. Existing feature representation learning-based approaches try to minimize the difference of latent features between different domains by exact alignment, which is achieved by either one-to-one topic alignment or matrix projection. Exact alignment, however, restricts the representation flexibility and further degrades the model performances on CLSC tasks if the distribution difference between two language domains is large. On the other hand, most previous studies proposed document-level models or ignored sentiment polarities of topics that might lead to insufficient learning of latent features. To solve the abovementioned problems, we propose a coarse alignment mechanism to enhance the model's representation by a group-to-group topic alignment into an aspect-level fine-grained model. First, we propose an unsupervised aspect, opinion, and sentiment unification model (AOS), which trimodels aspects, opinions, and sentiments of reviews from different domains and helps capture more accurate latent feature representation by a coarse alignment mechanism. To further boost AOS, we propose ps-AOS, a partial supervised AOS model, in which labeled source language data help minimize the difference of feature representations between two language domains with the help of logistics regression. Finally, an expectation–maximization framework with Gibbs sampling is then proposed to optimize our model. Extensive experiments on various multilingual product review data sets show that ps-AOS significantly outperforms various kinds of state-of-the-art baselines.**

***Index Terms*— Coarse alignment, cross-lingual sentiment classification (CLSC), topic model.**

I. Introduction

Consumers from different countries often write online reviews in different languages to express their opinions after buying products from Amazon or Alibaba, which are deemed valuable to producers, service providers, and consumers themselves. Generally, high-quality and annotated English review corpora are often available, whereas non-English corpora (e.g., Chinese and Japanese) are more difficult to obtain. Cross-lingual sentiment classification (CLSC) thus emerges as an important learning task and has attracted much attention from both academia and industries [1]–[3]. The key idea behind CLSC is to bridge the gap of vocabularies and/or semantics between the source and target language domains such that the resources in the source language domain can be adapted for target language domain. Along this line, machine-translation-based methods [4], transfer learning [5]–[7], and feature representation learning-based methods [8]–[11] have been proposed to solve cross-lingual problems.

Feature representation learning-based methods aim to induce a reasonable feature representation between the source and target language domains so as to reduce distributional differences. Many variants of topic models have been proposed to solve cross-lingual classification problems. For example, Lin *et al.* [9], Paul and Girju [12], Bao *et al.* [13], and Zhuang *et al.* [14] proposed to encode exact alignment by forcing domains to share the same common topics. Li *et al.* [15] utilized common topics to learn a projection matrix between different domains. However, the abovementioned methods have an intrinsic drawback, i.e., exact alignment of topics across domains. Specifically, the exact alignment is achieved by either one-to-one topic alignment or matrix projection. The abovementioned exact alignment restricts the representation flexibility and further depresses model performances when the distributional differences between the source and target language domains are large [3], that is, the assumption of exact alignment is often violated since different language domains usually differ in their underlying distributions. This motivates us to reduce the restrictions of *exact* alignments in modeling cross-lingual classification problems.

Manuscript received December 2, 2018; revised July 31, 2019, October 24, 2019, and January 20, 2020; accepted March 4, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1402800 and in part by the National Natural Science Foundation of China under Grant 71501003. The work of Junjie Wu was supported in part by the National Key R&D Program of China under Grant 2019YFB2101804 and in part by the National Natural Science Foundation of China under Grant 71531001, Grant 71725002, and Grant U1636210. The work of Fuzhen Zhuang was supported in part by the National Natural Science Foundation of China under Grant U1836206 and in part by the Project of Youth Innovation Promotion Association CAS under Grant 2017146. *(Corresponding author: Fuzhen Zhuang.)*

Deqing Wang, Chenwei Lu, and Chenguang Du are with the State Key Laboratory of Software Development Environment, School of Computer Science, Beihang University, Beijing 100191, China.

Baoyu Jing is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Junjie Wu and Guannan Liu are with the School of Economics and Management, Beihang University, Beijing 100191, China, and also with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China.

Fuzhen Zhuang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100864, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhuangfuzhen@ict.ac.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2979225

The second problem refers to the model granularity problem, i.e., which is more suitable for high-level representation learning, coarse-grained model, or fine-grained one? Generally, coarse-grained models only learn document-level feature representations, which often fail to capture various aspects (e.g., the screen, battery, and camera of an iPhone) in one real-life product review. As a matter of fact, fine-grained models [16], [17] outperform coarse-grained ones [18] for monolingual sentiment classification because the former can capture more accurate latent feature representations. However, for CLSC tasks, even though fine-grained model performs better than coarse-grained one in [9], it only achieves comparable performance to support vector machine (SVM) [10]. Similar to the model in [9], the abovementioned methods sample topics from source and target language domains by word-level translation, which might cause semantic drift and result in inaccurate topic-word distributions because of synonym and polysemy.

To address the abovementioned problems, we follow the theoretical work of [19] and try to learn a more accurate and flexible feature representation for CLSC. In this regard, we introduce a coarse alignment mechanism of topic and sentiment and then propose an aspects, opinions, and sentiments unification model named AOS, which not only distinguishes the common and specific topics across different language domains but also identifies sentiment polarity of topics and further coarsely aligns specific topics across different domains. AOS model helps learn a fine-grained representation for CLSC tasks. To further boost AOS, we propose an improved model named ps-AOS so that the labeled data in the source domain are used as partial supervision information to help minimize the distribution difference between the source and target domains, as well as the empirical loss in the source language domain. Finally, we present an EM framework with Gibbs sampling to infer the parameters of our model.

Our main research contributions are summarized as follows.

1) We propose an aspect-level unification model (AOS), which trimodels aspects, opinions, and sentiments of reviews from different domains and helps learn more accurate latent feature representation.
2) In unsupervised AOS, we introduce a coarse alignment mechanism to align specific topics with the same sentiment label of different domains, which overcomes the drawbacks of exact alignment in previous models.
3) To make full use of labeled instances in source language training data, we propose an improved model named ps-AOS with partial supervision, in which labeled source language data help minimize the difference of feature representations between the source and target language domains with the help of logistics regression.
4) We present an EM framework with Gibbs sampling to infer the parameters of our models and conduct extensive experiments to demonstrate the significant improvement of our models for cross-lingual and/or cross-domain sentiment classification tasks.

The remainder of this article is organized as follows. Section II reviews some related work on CLSC. Then, we formally present our model in Section III and give the experimental setup and results in Sections IV. Finally, Section V summarizes this article.

II. Related Work

A. Cross-Domain Adaptation

Cross-domain adaptation aims to extract the knowledge from the label-rich source domain to enhance the predictive model of the target domain. Existing methods often achieve knowledge transfer by detecting a shared low-dimensional feature representation from source domain to target domain. For example, Dai *et al.* [20] proposed a coclustering-based method, which identified the word clusters across different domains by propagating the class information and knowledge from source domain to target domain. Li *et al.* [21] proposed to share the same word clusters between the source and target domains to transfer label information. However, the word clusters between the source and target domains are only related, rather than the exactly same. Zhuang *et al.* [22] exploited the association between the word features concepts and the example classes as the bridge across domains.

Moreover, another recent studies argued that the high-level concepts help to model the difference of data distribution, and they are more appropriate for classification. Specifically, these methods assume that all the data domains have the same set of shared concepts or identical concepts, alike concepts, distinct concepts, which are used as the bridge for knowledge transfer. For instance, Wang *et al.* [23] first attempted to discover the alike concepts and used them for knowledge transfer. Then, Long *et al.* [5] divided shared concepts into identical and alike ones as cross-domain knowledge. Zhuang *et al.* [6], [24] exploited the identical, alike and distinct concepts for distinguishing knowledge. Hu *et al.* [25] proposed multiknowledge transfer from multisource domains to target domain.

To sum up, the abovementioned transfer learning approaches achieved a better performance than nontransferred methods for cross-domain sentiment classification. However, even though some approaches have noted the difference of knowledge or model transferred from source domain to target domain, they are two-stage transfer learning and could not consider transfer learning as a whole framework.

B. Cross-Lingual Adaptation

Compared with cross-domain adaptation, cross-lingual adaptation needs to address the nonoverlapped feature space problem. Thus, the key idea behind CLSC is to bridge the gap of vocabularies and/or semantics between the source and target language domains. Alone this line, many methods have been studied extensively and deeply [8], [10], [26]–[30]. For example, Banea *et al.* [31] leveraged a machine translation technique to improve the model performance of target language. Bilingual parallel corpora and dictionaries are also ideal resources for cross-language sentiment classification tasks [4]. For instance, Wan [32], [33] proposed a cotraining method, which applied bilingual reviews to improve the performance of classifier.

Topic model [34] is a widely used model for learning latent feature representation across different domains. Paul and Girju [12] proposed a cross-domain Latent Dirichlet Allocation model (ccLDA) and Bao *et al.* [13] combined partial supervision into ccLDA; these two models both encoded exact alignment by forcing specific topics between two domains to share the same topic indexes with common topics. Li *et al.* [15] utilized common topics to learn a projection matrix between specific topics of different domains. The disadvantage of the abovementioned methods lies in that exact alignment restricts the representation flexibility and results in a significant decline in accuracy when the distribution difference between the source and target language domains is large [3], that is, the assumption of exact alignment is often violated since different language domains usually differ in their underlying distributions. Moreover, the abovementioned models are coarse-grained ones because they only learn document-level feature representations, which often fail to capture various aspects in one real-life product review, e.g., the screen, battery, and camera of an iPhone. For fine-grained models, Lin *et al.* [9] sampled topics in word-level translation, but it might cause semantic drift and result in inaccurate topic-word distributions because of synonym and polysemy. Another challenge in a topic model is that exact inference is often intractable in topic models. Thus, some studies proposed to employ a stochastic EM framework [35], [36], which incorporated the functional optimization problem with Gibbs sampling [37].

Recently, deep neural networks (DNNs) have been applied to learn shared feature representations for cross-lingual sentiment analysis. For example, Chandar *et al.* [26] proposed a predicative autoencoder for learning shared representation. A compositional distributed semantics was learned in [38]. Jain and Batra [39] developed a cross-lingual sentiment analysis tool based on a bilingually constrained recursive autoencoder. Zhou *et al.* [27] proposed to learn bilingual word embedding for cross-lingual sentiment analysis. Generally, DNNs-based approaches need paired sentences from parallel corpora. Moreover, the learned feature representations are difficult to interpret and the algorithms own higher time complexities.

In this article, we first propose an aspect-level unification model (AOS), which trimodels aspects, opinions, and sentiments of reviews from different domains and helps learn more accurate latent feature representation by a coarse alignment mechanism. Then, to further boost AOS, we propose an improved model named ps-AOS, in which labeled source language data help minimize the difference of feature representations between the source and target language domains by applying logistics regression. Our proposed two methods achieve better performance than various state-of-the-art baselines on CLSC tasks.

III. Model and Inference

A. Generative Process of AOS and Ps-AOS

Given a source language domain $\mathcal{D}^s = \{(\mathbf{x}_1^s, p_1^s), \ldots, (\mathbf{x}_{N_s}^s, p_{N_s}^s)\}$ and a target language domain $\mathcal{D}^t = \{\mathbf{x}_1^t, \ldots, \mathbf{x}_{N_t}^t\}$, where N_s and N_t are the number of documents in source and target domains, respectively. $p \in \{+1, -1\}$ is the class label. First, we eliminate the gap of vocabularies between two language domains by machine translation, i.e., translating $\mathcal{D}^t$ into source language. For convenience, we use subscript to denote the translated domain, i.e., $\mathcal{D}_{trans}^t = \{\mathbf{x}_{trans_1}^t, \ldots, \mathbf{x}_{trans_{N_t}}^t\}$. Thus, predicting the sentiment label of each $\mathbf{x}_j^t$ in $\mathcal{D}^t$ becomes predicting the label of each $\mathbf{x}_{trans_j}^t$ by employing $\mathcal{D}^s$. Following the abovementioned models, our models have the following assumptions: 1) each sentence in reviews only belongs to one topic and has one sentiment polarity and 2) there are some domain-independent (common) and some domain-dependent (specific) topics across different domains.

In this article, we proposed an unsupervised model AOS and its improved model ps-AOS with partial supervision. Their graphical representations are shown in Fig. 1 together. It is worth noting that AOS is the plate without considering the orange shaded region, and its generative process is similar to ps-AOS without Step 3(e), that is, we do not draw a class label for document d in the source language domain by logistic regression. In order to present the generative process of our two models succinctly, we only show the generative process of ps-AOS as follows.

1) For each common topic z, the following steps hold.

   a) Choose $\phi_{l,z}^A \sim Dir(\beta)$ for each sentiment l.

2) For each domain c, the following steps hold.

   a) Choose $\phi_{c,l,z}^O \sim Dir(\beta)$ for common topic z and l.

   b) Choose $\psi_{c,l,z}^{A/O} \sim Dir(\beta)$ for specific topic z and l.

3) For each document d, the following steps hold.

   a) Choose a domain indicator c (not shown).

   b) Choose $\mu_d \sim Beta(\delta)$.

   c) Choose $\theta_{d,l} \sim Dir(\alpha)$, for each sentiment l.

   d) Choose $\sigma_{d,l} \sim Beta(\gamma)$, for each sentiment l.

   e) If d is from the source language domain, draw a class label $p_d \in \{-1, 1\} \sim Bern(logistic$ $(-p_d \eta^T \bar{\mathbf{z}}_d))$ *//ignore this step for AOS model.*

   f) For each sentence s, the following steps hold.

      i) Choose a sentiment $l_{d,s} \sim Bern(\mu_d)$.

      ii) Choose $r_{d,s} \sim Bern(\sigma_{d,l_{d,s}})$.

      iii) if $r_{d,s} = 0$ choose $z_{d,s} \sim Multi(\theta_{d,l_{d,s}}^I)$
      if $r_{d,s} = 1$ choose $z_{d,s} \sim Multi(\theta_{d,l_{d,s}}^S)$.

      iv) For each word n, the following steps hold.

         A) Choose $y_{d,s,n} \sim Bern(\pi_{d,s,n})$.

         B) If $r_{d,s} = 0$ and $y_{d,s,n} = 0$: $w_{d,s,n} \sim Multi$ $(\phi_{l_{d,s},z_{d,s}}^A)$
         if $r_{d,s} = 0$ and $y_{d,s,n} = 1$: $w_{d,s,n} \sim Multi$ $(\phi_{c,l_{d,s},z_{d,s}}^O)$
         if $r_{d,s} = 1$ and $y_{d,s,n} = 0$: $w_{d,s,n} \sim Multi$ $(\psi_{c,l_{d,s},z_{d,s}}^A)$
         if $r_{d,s} = 1$ and $y_{d,s,n} = 1$: $w_{d,s,n} \sim Multi$ $(\psi_{c,l_{d,s},z_{d,s}}^O)$.

The math notations are provided in Table I. Our ps-AOS model has three key components: 1) the fine-grained model helps capture fine-grained semantic information across two language domains; 2) sentiment variable l, which is crucial

TABLE I
MATH NOTATIONS

Notation	Description	Notation	Description
K^I/K^S	# of common/sepcific topics in total	C	# of domains
D	# of documents in a domain	S	# of sentences in a document
N	# of words in a sentence	ϕ^A/ψ^A	common/specific aspect-word distribution
L	# of sentiment classes	ϕ^O/ψ^O	common/specific opinion-word distribution
μ	the sentiment distribution for a document	θ^I/θ^S	common/specific topics for a document
σ	parameters for common and specific topic switching for a sentence	z	topic index for a sentence
π	parameters for aspect and opinion switching for a word	w	an observed word
p	class label for a document	l	the sentiment label for a sentence
x	feature vector for the maximum entropy (MaxEnt) model	λ	weights learned by the MaxEnt model
y	aspect and opinion switcher for a word	r	common and specific topic switcher for a sentence
α	Dirichlet prior parameter for θ	β	Dirichlet prior parameter for all word distributions
γ	symmetric Beta prior parameters for σ	δ	symmetric Beta prior parameters for μ
η	logistic regression coefficients	$\bar{\mathbf{z}}_d$	empirical topic distribution for document d
$Bern(\cdot)$	Bernoulli distribution with parameter$(\cdot)$	$Beta(\cdot)$	Beta distribution with parameter$(\cdot)$
$Multi(\cdot)$	Multinomial distribution with parameter$(\cdot)$	$Dir(\cdot)$	Dirichlet distribution with parameter$(\cdot)$

to both coarse alignment and fine graininess, helps reduce the distributional differences; and 3) partial supervision by embedded logistic regression enhances feature representation learning by explicitly minimizing the training error in the source language domain.

1) Fine Grained Topics: In our ps-AOS model, topics are first divided into domain-independent common topics ϕ and domain-dependent specific topics ψ. The switcher r is used to distinguish common and specific topics. Topics are then further fine-grained through sentiment variable l and aspect/opinions switcher y. Specifically, each common/specific topic ϕ_z/ψ_z is first divided into sentiment-associated subtopics $\phi_{l,z}/\psi_{l,z}$. Then, $\phi_{l,z}/\psi_{l,z}$ is subdivided into aspect $\phi^A_{l,z}/\psi^A_{l,z}$ and opinion $\phi^O_{l,z}/\psi^O_{l,z}$ parts. Finally, we assume that specific topics of one domain are independent of the others, and we have $\psi^A_{c,l,z}$, $\psi^O_{c,l,z}$ for each domain. For common topics, we assume that aspect parts are domain-independent, while we allow different domains to have different opinion words toward the same aspects; thus, we have $\phi^A_{l,z}$ and $\phi^O_{c,l,z}$.

2) Functionalities of Sentiment Variable: Sentiment variable has two roles, i.e., identifying sentiment of topics for fine graininess and aligning specific topics coarsely. For fine graininess, as shown in Fig. 1, we insert a sentiment variable l as topic z's parent to identify the sentiment polarity of z. Suppose that we have two reviews: Review_A is "I like the big screen of iPhone. But its price is too high," and Review_B is "I dislike the screen of iPhone, it is too huge for me." The first benefit of inserting sentiment variable l is that it allows aspects in the same topic to be associated with different sentiments. We can associate the aspect SCREEN with POSITIVE polarity from Review_A and the same aspect with NEGATIVE from Review_B. The second benefit is that subdividing one topic into different sentiment polarities will strengthen model's representation ability for classification tasks. In our model, <SCREEN, POSITIVE> and <SCREEN, NEGATIVE> are considered as two different features, whereas they are represented as one feature <SCREEN> in previous models. We validate the effectiveness of this fine graininess through extensive experiments.

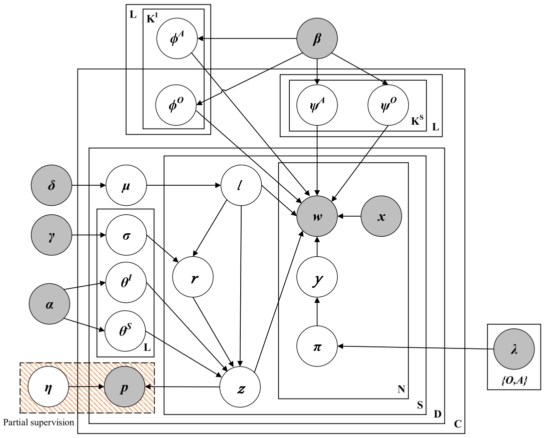


Fig. 1. Plate notation of the proposed AOS and ps-AOS model, where AOS is the plate without the orange shaded region. (Best viewed in color).

3) Coarse Alignment: The major issue for existing cross-lingual topic models is that they try to learn an exact alignment for topics in different domains, either through one-to-one alignment (CLJAS [9], ccLDA [12], and PSCCLDA [13]) or projection matrix (TCA [15]). To address this problem, we introduce a coarse alignment mechanism, which is an alignment between topic groups. For briefness, we define a topic group as a set of topics sharing the same sentiment. The coarse alignment is encoded in the generative process by assuming that a topic $z_{d,s}$ and its common/specific switcher $r_{d,s}$ are generated after sentiment $l_{d,s}$ has been chosen. Under such a generative process, specific topics with the POSITIVE label will forcibly be aligned with common topics with the POSITIVE label. Thus, specific topics with POSITIVE labels in different domains will be aligned via common topics with POSITIVE labels. The topics with NEGATIVE labels are similarly aligned.

4) Partial Supervision: To further improve AOS, we introduce partial supervision into AOS and propose its variant ps-AOS. Here, "partial supervision" means that the model only observes the class labels in the source language domain. To reduce the training error, we adopt class labels in the source domain to guide the sampling of topics by partial supervision in ps-AOS, which will enhance learned feature representations. Specifically, we draw a class label p_d for each review in source domain from $Bern(logistic(-p_d\eta^T\bar{\mathbf{z}}_d))$ in Step 3(e) while doing nothing to reviews in the target domain. Then, the topic $z_{d,s}$ of source domain is sampled according to 1. We can observe that the sampling of $z_{d,s}$ is related to p_d and η. Actually, our model considers class labels of examples in the source domain into the generative process of topics by parameters p_d and η. Under the partial supervision, for a sentence s in d from source domain, if the label of d is NEGATIVE, then the probability of $z_{d,s}$ belonging to NEGATIVE sentiment increases. With the supervision for generating topics, feature representations will be enhanced by reducing training error.

Therefore, according to the theoretical work in [19], a good representation enables achieving a low error rate in the source domain as well as minimizing a distance between the induced marginal distributions of the two domains. In our model, the embedded logistic regression explicitly models class labels of examples in the source domain, which could help minimize the empirical training error in the source domain and 2) new representation learned by aspect-level fine-grained model and coarse alignment can achieve lower proxy $\mathcal{A}$-distance (PAD) value, which indicates that the new representations of both source and target domains are as indistinguishable as possible [40]. We will provide experimental verification of PAD in Section IV.

B. Inference

Exact posterior inference is intractable in AOS and ps-AOS models, so we employ a collapsed Gibbs sampling algorithm [37] for approximate inference. To solve the optimization problem of our proposed AOS and ps-AOS models, we propose an EM framework. In E-steps, we fix η and use collapsed Gibbs sampling to infer hidden variables (z, r, l, and y) and then update the empirical topic representation $\bar{\mathbf{z}}_d$ of document d. In M-steps, we update the logistic regression coefficient η by maximizing the joint likelihood of training examples in the source domain.

Note that the maximum entropy (MaxEnt) component is trained before performing Gibbs sampling, which means $\{\lambda_O, \lambda_A\}$[1] are fixed during Gibbs sampling. Thus, in E-step, we have four sets of latent variables for ps-AOS model: z, r, l, and y and can jointly sample $(z_{d,s}, r_{d,s}, l_{d,s})$ as a block in 1, given the assignments of all other hidden variables, that is

$$
\begin{aligned}
&P(z_{d,s}=k, r_{d,s}=j, l_{d,s}=t \mid \mathbf{z}_{\neg(d,s)}, \mathbf{r}_{\neg(d,s)}, \mathbf{l}_{\neg(d,s)}, \mathbf{y}, \mathbf{w}, \mathbf{x})\\
&\propto \frac{C^{d}_{(t)}+\delta}{C^{d}_{(\cdot)}+L\delta} \times \frac{C^{d,t}_{(j)}+\gamma}{C^{d}_{(t)}+2\gamma} \times \frac{C^{d,t,j}_{(k)}+\alpha}{C^{d,t}_{(j)}+K^{j}\alpha}\\
&\times \frac{1+e^{-p_d\eta^T\bar{\mathbf{z}}_d}\cdot e^{p_d\eta_k/C^{d}_{(\cdot)}}}{1+e^{-p_d\eta^T\bar{\mathbf{z}}_d}}\\
&\times \left(\frac{\Gamma\left(C^{A,t,j,k}_{(\cdot)}+V\beta\right)}{\Gamma\left(C^{A,t,j,k}_{(\cdot)}+N^{A,t,j,k}_{(\cdot)}+V\beta\right)} \cdot \prod_{v=1}^{V}\frac{\Gamma\left(C^{A,t,j,k}_{(v)}+N^{A,t,j,k}_{(v)}+\beta\right)}{\Gamma\left(C^{A,t,j,k}_{(v)}+\beta\right)}\right)\\
&\times \left(\frac{\Gamma\left(C^{O,t,j,k}_{(\cdot)}+V\beta\right)}{\Gamma\left(C^{O,t,j,k}_{(\cdot)}+N^{O,t,j,k}_{(\cdot)}+V\beta\right)} \cdot \prod_{v=1}^{V}\frac{\Gamma\left(C^{O,t,j,k}_{(v)}+N^{O,t,j,k}_{(v)}+\beta\right)}{\Gamma\left(C^{O,t,j,k}_{(v)}+\beta\right)}\right)
\end{aligned}
\tag{1}
$$

where $C^{d}_{(t)}$ is the number of sentences assigned with label t in document d, $C^{d,t}_{(j)}$ is the number of sentences assigned to type j (common/specific) topics with sentiment label t in document d, $C^{d,t,j}_{(k)}$ is the number of sentences assigned to topic k of type j with sentiment label t, $C^{A,t,j,k}_{(v)}$ and $C^{O,t,j,k}_{(v)}$ are the number of times word v is assigned as an aspect and opinion word to topic k of type j with sentiment t, respectively. $N^{A,t,j,k}_{(v)}$ and $N^{O,t,j,k}_{(v)}$ are the number of times word v is assigned as an aspect and opinion word to topic k of type j with sentiment t in sentence s of document d, respectively, $C^{x}_{(\cdot)}$ is the total number of sentences assigned with x, and $N^{x}_{(\cdot)}$ is the total number of words assigned with x in sentence s of document d. Note that all these counts represented by symbol $C(\cdot)$ exclude sentence s of document d.

For AOS model, in E-step, we only have three sets of latent variables: z, r, and l, i.e., without considering y. Thus, the inference for triplet (z, r, l) in AOS model is as follows:

$$
\begin{aligned}
&P(z_{d,s}=k, r_{d,s}=j, l_{d,s}=t \mid \mathbf{z}_{\neg(d,s)}, \mathbf{r}_{\neg(d,s)}, \mathbf{l}_{\neg(d,s)}, \mathbf{y}, \mathbf{w}, \mathbf{x})\\
&\propto \frac{C^{d}_{(t)}+\delta}{C^{d}_{(\cdot)}+L\delta} \times \frac{C^{d,t}_{(j)}+\gamma}{C^{d}_{(t)}+2\gamma} \times \frac{C^{d,t,j}_{(k)}+\alpha}{C^{d,t}_{(j)}+K^{j}\alpha}\\
&\times \left(\frac{\Gamma\left(C^{A,t,j,k}_{(\cdot)}+V\beta\right)}{\Gamma\left(C^{A,t,j,k}_{(\cdot)}+N^{A,t,j,k}_{(\cdot)}+V\beta\right)} \cdot \prod_{v=1}^{V}\frac{\Gamma\left(C^{A,t,j,k}_{(v)}+N^{A,t,j,k}_{(v)}+\beta\right)}{\Gamma\left(C^{A,t,j,k}_{(v)}+\beta\right)}\right)\\
&\times \left(\frac{\Gamma\left(C^{O,t,j,k}_{(\cdot)}+V\beta\right)}{\Gamma\left(C^{O,t,j,k}_{(\cdot)}+N^{O,t,j,k}_{(\cdot)}+V\beta\right)} \cdot \prod_{v=1}^{V}\frac{\Gamma\left(C^{O,t,j,k}_{(v)}+N^{O,t,j,k}_{(v)}+\beta\right)}{\Gamma\left(C^{O,t,j,k}_{(v)}+\beta\right)}\right).
\end{aligned}
\tag{2}
$$

With assignments of $\mathbf{z}$, $\mathbf{r}$, and $\mathbf{l}$, we can sample $y_{(d,s,n)}$ for $y_{(d,s,n)}=0$

$$
\begin{aligned}
&p(y_{(d,s,n)}=0 \mid \mathbf{z}, \mathbf{r}, \mathbf{l}, \mathbf{y}_{\neg(d,s,n)}, \mathbf{w}, \mathbf{x})\\
&\qquad\propto \frac{\exp(\lambda_A\cdot\mathbf{x}_{d,s,n})}{\Sigma_{i\in\{O,A\}}\exp(\lambda_i\cdot\mathbf{x}_{d,s,n})} \times \frac{C^{A,l_{d,s},r_{d,s},z_{d,s}}_{(w_{d,s,n})}+\beta}{C^{A,l_{d,s},r_{d,s},z_{d,s}}_{(\cdot)}+V\beta}
\end{aligned}
$$

[1] Similar to MaxEnt-LDA [41], we adopt a maximum entropy model trained with part-of-speech (POS) tags of words to compute hyperparameter λ for π.

and for $y_{(d,s,n)} = 1$

$$p(y_{(d,s,n)} = 1|\mathbf{z}, \mathbf{r}, \mathbf{l}, \mathbf{y}_{\neg(d,s,n)}, \mathbf{w}, \mathbf{x}) \propto \frac{\exp(\lambda_O \cdot \mathbf{x}_{d,s,n})}{\Sigma_{i\in\{O,A\}}\exp(\lambda_i \cdot \mathbf{x}_{d,s,n})} \times \frac{C^{O,l_{d,s},r_{d,s},z_{d,s}}_{(w_{d,s,n})} + \beta}{C^{O,l_{d,s},r_{d,s},z_{d,s}}_{(\cdot)} + V\beta}$$

where $\mathbf{x}_{d,s,n}$ is the POS feature vector.

With the abovementioned assignments, the approximate probability (μ) of sentiment l in d, the approximate probability (θ) of topic k of type j with sentiment polarity l in d, and the approximate probabilities of word w in topic k of type $j = 0$ (ϕ) and $j = 1$ (ψ) with sentiment l can be estimated as follows:

$$\mu^d_{(t)} = \frac{C^d_{(t)} + \delta}{C^d_{(\cdot)} + L\delta}, \quad \theta^{d,t,j}_{(k)} = \frac{C^{d,t,j}_{(k)} + \alpha}{C^{d,t,j}_{(\cdot)} + K^j\alpha}$$

$$\phi^{t,k}_{(v)} = \frac{C^{t,j=0,k}_{(v)} + \beta}{C^{t,j=0,k}_{(\cdot)} + K^I\beta}, \quad \psi^{t,k}_{(v)} = \frac{C^{t,j=1,k}_{(v)} + \beta}{C^{t,j=1,k}_{(\cdot)} + K^S\beta}.$$

After obtaining the new empirical topic representation $\bar{\mathbf{z}}_d$ of each document d in the source training domain, in M-step, we update the logistic regression coefficient η through $\bar{\mathbf{z}}_d$. Then, the updated η is used in the E-step of the next iteration.

Note that the computational complexity of our proposed AOS and ps-AOS models is similar to the traditional topic model. Concretely, for each sentence, we need to sample z, r, and l, and the time complexity of the abovementioned sampling steps can be seen as $O(1)$. Therefore, the computational complexity of each iteration in our proposed models is $O((K + L + 4)N)$, where K is the number of topics, N is the number of sentences, and L is the number of sentiment labels (e.g., $L = 2$ for binary sentiment classification). For traditional LDA, its computational complexity of each iteration is $O(KN)$ [42]. The other variants based on LDA have the similar computational complexity, such as TSU [16] and PSCCLDA [13].

C. Cross-Lingual Sentiment Classification

After applying the EM algorithm with Gibbs sampling described in Section III-B, we obtain the topic distribution of each review and use them to represent reviews in high-level latent feature space. Specifically, $\bar{\mathbf{z}}_d$ is an $L * (K^I + K^S)$-dimensional vector, in which the coarsely aligned specific topics in different domains have the same topic indexes. In our product review classification, the review is labeled as positive or negative, so we set $L = 2$ in our experiments; and we predict their sentiment labels according to the equation $p(p_d|\bar{\mathbf{z}}_d) = 1/(1 + e^{-p_d\eta^\top\bar{\mathbf{z}}_d})$. If $p(p_d = 1|\bar{\mathbf{z}}_d) \geq 0.5$, the predicted label of document d is positive and negative otherwise.

IV. Experiments

A. Data Sets

Multilingual Amazon product reviews data set[2] comprises of three product categories: BOOKS (B), DVD (D), and MUSIC (M). It has reviews in four languages: English (E), German (G), French (F), and Japanese (J). For each category in each language, the numbers of both training data and test data are 2000, i.e., 1000 positive and 1000 negative examples. The second one is the balanced English-Chinese NLP&CC cross-language data set.[3] It also contains the abovementioned three categories. However, there are 2000 positive and 2000 negative reviews. In our experiments, we use English as the source language and each of the other four languages as the target language. Note that the two data sets also provide the corresponding English test reviews of each target language test reviews, which are translated by Google Translate. Therefore, we directly use the translated reviews to predict their categories in AOS and ps-AOS models. Finally, we obtain 12 CLSC tasks and 18 cross-lingual/domain tasks, as shown in Table II. We perform the same data preprocessing for all of the models: remove punctuations, stop words and nonalpha characters.

B. Baseline Methods and Setup

We compare our proposed method with various kinds of state-of-the-art baseline algorithms. They are given in the following.

1) *Machine-Translation-Based Methods (MT-SVM with linear kernel and MT-LR) [4]):* We obtained the best values of their parameters by fine-tuning on EFB task.
2) *Transfer Learning Methods (DTL [5] and TRiTL [6]):* For DTL and TRiTL, we first fine-tuned their parameters on the EFB task and then adopted the fine-tuned values as the default parameters of the rest tasks in order to raise efficiency of our experiments. Finally, the parameters of DTL and TRiTL are as follows. The numbers of iterations of both DTL and TRiTL are set as 180, the numbers of common/specific topics in DTL are 25/25, and the numbers of identical concepts, alike concepts, and distinct concepts in TRiTL are set as 20, 20, and 10, respectively.
3) *Semi-Supervised Methods (CL-SCL [43] and Co-training [33]):* Note that CL-SCL and cotraining need auxiliary data. After fine-tuning on the EFB task, for CL-SCL, we set $m = 600$ and $k = 120$ for a better performance, while the rest parameters are the same with CL-SCL (i.e., $\phi = 30$ and $\lambda = 10^{-5}$), which achieves the best performance. For cotraining, we adopted two monolingual SVM to select ten most confidently predicted instance (five positive instances and five negative ones).
4) *Topic Models on Singular Domain (TSU [16]):* After tuning on the EFB task, the parameters for TSU are: #topics = 15, #iterations = 1000, $\alpha = \beta = \eta = 0.01$, and $\gamma = 1$. JST and ASUM are excluded because TSU outperforms them significantly.
5) *Cross-Domain Topic Models (TCA [15] and PSC-CLDA [13]):* For the two models, we also fine-tuned them on the EFB task and then set their parameters

[2] http://www.uni-weimar.de/medien/webis/corpora/corpus-webis-cls-10/cls-acl10-processed.tar.gz

[3] http://tcci.ccf.org.cn/conference/2013/dldoc/evdata03.zip

TABLE II

CLASSIFICATION ACCURACIES (%) FOR CROSS-LINGUAL AND CROSS-DOMAIN SENTIMENT CLASSIFICATION TASKS

Task	LR	SVM	TRiTL	DTL	CL-SCL	Co-training	TSU	TCA	PSCCLDA	AOS(exact)	AOS	ps-AOS
ECB	65.8	70.1	65.3	65.6	78.1	81.4	80.4	70.0	67.7	86.4	88.7	**90.3**
ECD	65.8	65.2	65.5	65.0	78.1	83.1	81.3	68.4	61.0	86.4	89.2	**90.1**
ECM	62.8	63.3	67.3	62.7	74.9	77.2	80.8	69.5	60.8	86.8	87.4	**90.2**
EFB	80.7	78.4	81.1	81.4	82.3	81.5	72.6	60.5	76.4	73.4	84.0	**84.6**
EFD	78.7	76.5	80.8	82.1	79.5	81.6	74.8	56.3	62.9	75.3	**82.6**	82.0
EFM	79.1	74.7	77.5	78.3	77.4	80.3	76.7	53.5	64.5	79.4	83.0	**84.9**
EGB	81.3	78.9	81.7	81.9	83.1	81.5	71.9	61.2	62.7	75.0	84.1	**84.9**
EGD	79.3	76.3	78.2	79.7	82.5	79.7	71.2	58.8	63.5	76.8	83.4	**84.5**
EGM	79.3	75.9	80.3	81.3	82.0	80.2	71.9	57.5	64.6	72.4	83.9	**85.8**
EJB	70.6	68.5	71.5	71.8	73.8	76.7	67.7	52.4	61.4	72.6	81.0	**83.5**
EJD	73.5	70.1	71.6	72.5	75.5	82.1	70.9	55.5	61.2	72.8	81.9	**82.9**
EJM	73.5	70.9	76.1	77.5	78.9	81.4	72.0	60.7	60.4	73.9	85.4	**86.0**
Avg.	74.2	72.4	74.7	75.0	78.8	80.6	74.4	60.4	63.9	77.6	84.6	**85.8**
EBFD	80.0	77.3	78.0	80.7	76.8	79.3	74.7	56.0	60.8	78.8	84.0	**85.8**
EBFM	76.3	74.3	78.0	79.3	77.7	79.1	77.0	57.4	61.6	78.8	84.6	**85.7**
EBGD	77.4	75.4	77.0	75.6	79.2	75.4	74.1	54.3	58.1	71.1	81.7	**86.2**
EBGM	75.9	73.6	78.1	79.2	81.3	81.6	71.8	52.2	62.1	74.1	81.8	**85.2**
EBJD	71.4	70.1	70.4	70.1	75.8	81.2	72.9	52.0	58.1	73.7	81.8	**84.3**
EBJM	69.5	67.6	74.8	76.9	74.1	82.5	73.9	51.8	57.4	78.4	82.6	**83.2**
EDFB	76.4	74.8	80.2	79.9	80.9	75.1	73.6	58.4	76.2	75.8	**84.6**	83.9
EDFM	77.8	74.7	79.4	79.3	77.4	79.3	77.0	59.4	63.9	77.2	81.7	**85.5**
EDGB	78.8	76.3	81.0	80.9	83.9	79.8	71.3	57.0	63.1	74.2	80.2	**85.9**
EDGM	77.8	76.0	78.2	78.7	82.4	83.4	71.8	51.7	60.1	72.9	83.0	**84.6**
EDJB	69.9	68.4	69.6	71.0	74.9	74.0	73.9	55.6	58.9	72.8	81.8	**83.8**
EDJM	73.8	71.5	73.3	78.3	77.8	82.7	71.1	52.2	59.7	75.4	81.9	**84.7**
EMFB	76.7	74.6	77.9	77.2	80.1	82.5	73.6	53.0	72.9	75.2	**83.5**	82.4
EMFD	79.7	77.4	80.4	78.5	78.8	79.0	74.7	53.9	61.3	76.7	84.2	**84.7**
EMGB	76.8	74.2	79.4	77.7	83.4	79.9	71.3	55.4	60.8	73.0	81.5	**85.8**
EMGD	77.5	75.7	77.8	77.2	79.3	81.6	74.1	53.9	63.4	72.4	82.5	**85.5**
EMJB	68.4	67.1	68.1	67.2	69.4	76.5	71.1	50.3	60.2	69.3	**81.7**	81.6
EMJD	70.4	68.6	72.5	70.9	76.2	81.2	72.9	52.3	58.5	73.8	82.3	**82.7**
Avg.	75.3	73.2	76.3	76.6	78.3	79.7	73.4	54.3	62.1	74.6	82.5	**84.5**

Note: EFB denotes that English BOOKS data is source domain and French BOOKS data is target domain;
EBFD denotes that English BOOKS data is source domain and French DVD data is target domain.

as follows: #iterations $= 1000$ and #common/specific topics is 50/50. Besides, for PSCCLDA, $\alpha = \beta = 0.01$.

6) *Our Proposed Baseline (AOS(exact))*[4]*:* We propose a baseline to compare the effectiveness of coarse alignment mechanism. AOS(exact) implements an one-to-one alignment of topics between two domains by assuming that topic type switcher r and sentiment variable l are generated after topic z has been selected. The hyperparameters of our proposed models are set as follows: #iterations $= 300$, $K^I = K^S = 10$, $\alpha = \beta = \gamma = 0.01$, and $\delta = 1e^{-4}$.

For the initialization of models containing sentiment variables, we use MPQA subjectivity lexicon[5] to determine the initial sentiment assignments for words/sentences, as done in [9]. For a word, its sentiment label is initialized by its sentiment polarity. For a sentence, we first find all of sentiment words in it and then use simple majority voting to determine its initial sentiment assignment.

For the baselines, LR, SVM, TRiTL, and DTL belong to supervised models because they use the labeled data in the source domain. CL-SCL and cotraining are semisupervised models because they employ auxiliary data except for training data. TSU, TCA, AOS(exact), and AOS are unsupervised models because they only use training and test data to learn high-level features without using the label information of source language data, and thus, a logistic regression classifier is trained on high-level features of reviews in the source domain and is employed to predict the labels of reviews in the target domain for all the unsupervised models. Compared with the abovementioned unsupervised topic models, PSCCLDA and ps-AOS are partially supervised models because they consider the label information of source language data when learning high-level features.

C. Classification Performance

It is obvious that ps-AOS is not only suitable for CLSC tasks but also suitable for cross-domain sentiment classifications. In this section, we show both cross-lingual and cross-domain sentiment classification experimental results. The classification accuracies are presented in Table II. According to the Wilcoxon signed-rank test [44] (z-score $=$ -4.782 l), our ps-AOS significantly outperforms the second best method (AOS) on 30 kinds of cross-lingual and cross-domain tasks.

On average, for CLSC tasks, the accuracies of AOS and ps-AOS are improved up to 4% and 5.2%, respectively, compared with the second best cotraining. For cross-domain tasks, the improvements are 2.8% and 4.8%, respectively.

The ps-AOS model improves the averaged accuracy over traditional LR and SVM by at least 7%. Meanwhile, transfer learning methods (DTL and TRiTL) obtain better performance

[4] Model and inference can be found in the Appendix.

[5] http://mpqa.cs.pitt.edu/lexicons

TABLE III

TOP FIVE WORDS OF ASPECTS WITH CORRESPONDING OPINIONS AND SENTIMENTS ON ENGLISH AND CHINESE DVD REVIEWS

common topic (source domain)				specific topic (source domain)			
NEGATIVE		POSITIVE		NEGATIVE		POSITIVE	
Aspect	Opinion	Aspect	Opinion	Aspect	Opinion	Aspect	Opinion
movie	poor	movie	great	movie	bad	book	great
stars	bad	people	good	people	killing	story	work
quality	disappointed	story	beautiful	character	wrong	life	young
picture	boring	time	enjoy	man	problem	read	american
people	funny	series	perfect	time	waste	time	fascinating

common topic (translated target domain)				specific topic (translated target domain)			
NEGATIVE		POSITIVE		NEGATIVE		POSITIVE	
Aspect	Opinion	Aspect	Opinion	Aspect	Opinion	Aspect	Opinion
movie(电影)	bad(坏的)	movie(电影)	good(好的)	quality(质量)	poor(差劲的)	quality(质量)	good(好的)
stars(星级)	poor(差劲的)	people(人物)	classic(经典的)	price (价格)	expensive(贵的)	picture(画面)	clear(清楚的)
quality(质量)	stuck(卡住的)	story(故事)	clear(清楚的)	money(钱)	bad(坏的)	content(内容)	worthy(值得的)
picture(画面)	waste(浪费)	time (时代)	funny(有趣的)	content(内容)	regretful(后悔的)	disc(唱片)	pretty(优美的)
people(人物)	garbage(垃圾)	series(系列)	worthy(值得的)	book(书籍)	broken(损坏的)	movie(电影)	classic(合理的)

than LR and SVM, but their accuracies are lower about 6% than our model. CL-SCL utilizes many auxiliary unlabeled target language examples, so its accuracy is higher than machine-translation methods and transfer learning. However, CL-SCL is lower than cotraining and our method in terms of accuracy. Our ps-AOS model only takes as input the source language examples and the translated target language ones. We can observe that comodeling two domains helps improve classification accuracy significantly from the results shown in Table II. Compared to cotraining, our ps-AOS model improves about 5% in terms of accuracy because it provides an informative high-level latent feature representation for reviews, rather than noisy raw word representation.

For unsupervised topic models, TSU performs much better than TCA and PSCCLDA because TSU is a fine-grained model and it can capture more accurate latent features. It also shows that fine-grained model performs better for sentiment classification task. Therefore, we also adopt fine-grained model for cross-domain situations, where topics are divided into common and specific topics and each topic is further subdivided by sentiment, aspect, and opinion. For classification, each topic with different sentiment polarities is considered as different latent feature representations of reviews. Due to the representation flexibility introduced by coarse alignment and partial supervision introduced by logistic regression, the accuracy is improved about 11% by ps-AOS, compared with TSU. For cross-domain models, TCA and PSCCLDA achieve the worst performance in general. This is because TCA and PSCCLDA belong to a coarse-grained document-level model and thus fail to capture the aspects and sentiment details of each review and the exact alignment restricts the representation ability of the models and results in worse accuracy when the source and target domains have different distributions.

To verify the effectiveness of the coarse alignment mechanism and partial supervision, we compare ps-AOS with our proposed baseline models [AOS and AOS(exact)] under the same parameter settings. The only difference between AOS and AOS(exact) is the topic alignment mechanism: coarse alignment versus exact alignment. As shown in Table II, we can observe that coarse alignment improves the averaged accuracy over AOS(exact) by 7%. Compared with AOS, the partial supervision (ps-AOS) further improves the classification accuracy by 1.2% and 2% for cross-lingual and cross-domain tasks, respectively,

From the abovementioned experiments, we can observe that our ps-AOS not only constructs a fine-grained model to help capture more accurate feature representations across different domains but also employs coarse alignment and partial supervision to help minimize the differences of latent feature representations across domains for CLSC tasks.

D. Mining Aspects, Opinions, and Sentiments

Besides classification, ps-AOS can mine aspects, opinions, and sentiment polarities, simultaneously. Table III shows the top five words of randomly selected one common topic and two specific topics with different sentiment polarity from the English and Chinese DVD reviews data sets. Note that the indexes of two specific topics are the same.

First, we can observe that topics with different polarities focus on different aspects. For example, the common aspect with negative sentiment focuses on "quality" and "picture" of DVD products, whereas the common aspect with positive sentiment is about "people" and "story." Second, for common topics in different language domains, we find that the opinion words of the same aspect are different. For example, Chinese consumers express their negative opinions by using "stuck" and "garbage," which do not occur in English consumers' reviews. Third, the content differences are more obvious for specific topics. For instance, English customers mainly complain "people" or "character" of DVD products in the specific aspect with negative sentiment, whereas Chinese consumers mainly express their dissatisfaction about "quality" or "price." For example, we retrieve many sentences, such as "the dvd is broken" or "the price is expensive" from Chinese reviews. For positive sentiment, English customers praise for "story" or "books" of DVD, whereas Chinese consumers focus more on "picture" or "disc." Through the earlier observations, we can find the mined aspects and opinions with different sentiment polarities indeed reflect customers' different focuses on the same products and their opinion expression habits.

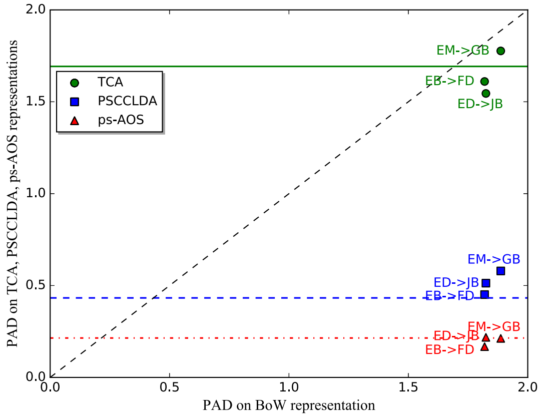


Fig. 2. PAD values on different feature representations.

The aspects, opinions, and sentiments mined by our model not only reveal the similarity and difference between different language consumers' reviews on the same products but also distinguish aspects with different sentiment as different latent features to strengthen the feature representation ability.

E. Proxy Distance of Feature Representation

Ben-David's theoretical work shows that PAD is a metric estimating the similarity of the source and target representations [19]. Generally, a lower PAD value indicates a better representation between two domains. We obtain the PAD value as done in [40] and verify that feature representations learned by ps-AOS in the source and target domains are difficult to distinguish. We calculate the averaged PAD values of 18 cross-lingual and cross-domain tasks and compare ps-AOS with other cross-domain models (TCA [15] and PSCCLDA [13]).

According to the Wilcoxon signed-rank test [44], our ps-AOS significantly outperforms TCA and PSCCLDA in terms of PAD value on the 18 tasks at a significance level of 0.01 (p-values are 0.0004 and 0.0002, respectively). As shown in Fig. 2, the averaged PAD values (three horizontal lines) of ps-AOS, PSCCLDA, and TCA are 0.215, 0.433, and 1.693, respectively. To illustrate it clearly, Fig. 2 also shows the PAD values of different algorithms on randomly selected three tasks (EBFD, EDJB, and EMGB). We can observe that TCA, PSCCLDA, and ps-AOS all achieve lower PAD values than bag-of-words (BoW) representation. It means that the learned high-level latent feature representation helps minimize the semantic gap between the source and target domains. Compared with TCA and PSCCLDA, ps-AOS further achieves lower PAD values. These observations also explain that sentiment variable and partial supervision in ps-AOS both help learn a better feature representation between two language domains, which plays a critical role on the improvement of classification accuracy in the target language domain.

F. Parameters Analysis

In ps-AOS, there are four hyperparameters. Fig. 3(a) and (b) shows the influence of hyperparameters on accuracy for EFB and EDGM tasks. First, we can observe that the classification accuracy increases as the values of α and β increase. This indicates, to some degree, that larger α and β better approximate the true distributions of the data sets. Actually, larger α and β imply that the latent semantic of a review comprises more topics, and words in each topic have more co-occurrences [45]. This is because larger α and β, hyperparameters for symmetric Dirichlet distributions, encode stronger assumption that θ and ϕ/ψ are uniformly distributed. From Fig. 3(a) and (b), we observe that $\alpha = 0.01$ can achieve the best accuracy for both EFB and EDGM tasks. However, the best values of β for both EFB and EDGM tasks are 0.001 and 0.01, respectively.

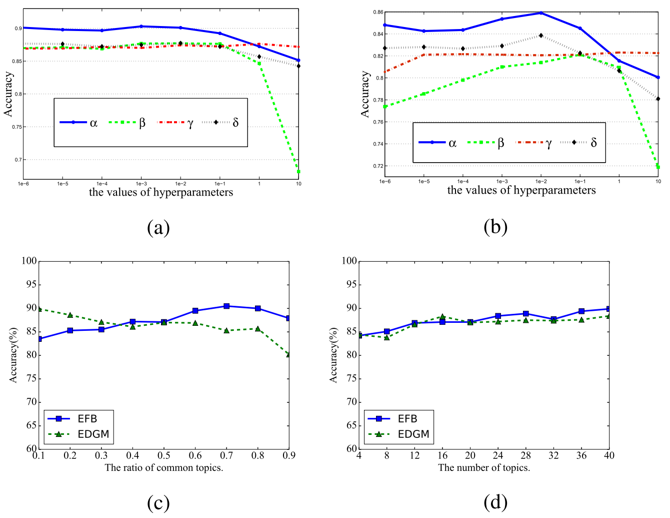


Fig. 3. Parameters analysis. (a) EFB task. (b) EDGM task. (c) Ratio comparison. (d) Total topics.

In addition, δ is the parameter of symmetric Beta prior distribution for μ, which is used to approximate the true sentimental distribution of sentences in each review, i.e., the proportions of POSITIVE/NEGATIVE sentiments in each review. Similarly, γ is the parameter of symmetric Beta prior distribution for σ, which is used to approximate the true topic distribution of common/specific topics in each review, i.e., the proportions of common/specific topics in each review. Generally, in each review d, μ_d is review-dependent and is determined by the proportions of POSITIVE/NEGATIVE sentences in d, and σ_d is also review-dependent and is determined by the proportions of common/specific topics in d. For example, for a review with a POSITIVE label, the proportion of positive sentences is larger than that of negative ones, and thus, smaller δ could fit the abovementioned situation better. As shown in Fig. 3(c), we can observe that the model achieves the best performance when the ratio of common topics is set to 0.9 in the same domain task (EFB) and 0.1 in cross-domain task (EDGM). The earlier observation indicates that the proportions of common/specific topics in each review should be uneven, and thus, larger γ in the EFB task [Fig. 3(a)] and smaller γ in the EDGM task [Fig. 3(b)] could fit the distribution of common/specific topics better. Therefore, for EFB task, the best values of δ and γ are $1e^{-4}$ and 1, whereas the best values for EDGM task are 0.01 and $1e^{-5}$.

In addition to the four hyperparameters, the ratio of common topics and the total number of topics will also influence the performance of the model. Fig. 3(c) shows the trend of accuracy when the ratio is changed from 0.1 to 0.9. In this experiment, the total number of topics is fixed as 20. It can be observed that our model performs best when the ratio of common topics falls in [0.6, 0.9] for EFB task. The potential reason behind this might be that the source and target domains are both about BOOK despite the language differences, and thus, it is reasonable that two domains share many common topics. For the EDGM task, the model achieves the best accuracy when the ratio of common topics is 0.1. We believe that this is due to the fact that source and target domains of EDGM task come from DVD and MUSIC, respectively, and thus, it is reasonable for these two domains to share only a small number of common topics. Fig. 3(d) shows the trend of accuracy when the total number of topics is changed from 4 to 40. Generally, the accuracy increases as the total number of topics increases. The potential reason is that the model may capture more semantic details about the reviews as the number of topics increase.

V. Conclusion

In this article, we jointly model aspects, opinions, and sentiments through a coarse alignment in a partially supervised way for CLSC tasks. Through the proposed ps-AOS model, we can mine polarized cross-lingual topics and their corresponding opinions in a coarse alignment manner. Moreover, we adopt logistic regression to make full use of labeled data in the source domain to minimize the difference of latent feature representations between two domains. Experimental results demonstrate the effectiveness of our model over various kinds of baselines on CLSC tasks.

However, our proposed AOS and ps-AOS models still have some limitations to be improved. For example, the number of common/specific topics across domains is predefined and data-dependent. In the future, we need some algorithms to help determine the number of common/specific from the training data automatically. Another improved direction is how to extend our model to solve multiclass cross-domain text classification problems, in which we need to replace the Bernoulli distribution of class labels with multinomial distribution.

Appendix
The Model of AOS(Exact)

In the appendix, we show the plate notation of AOS(exact) model and its corresponding inference results for triplet (z, r, l). Similar to the ps-AOS model, the math notations used are provided in Table I, Fig. 4 shows the plate notation of AOS(exact), and the blue dashed arrows show the differences between AOS and AOS(exact). AOS(exact) assumes that topic z is generated first, and then, common/specific topic switcher r and sentiment variable l are generated. Thus, it forces the common and specific topics that share the same topic index to be aligned.

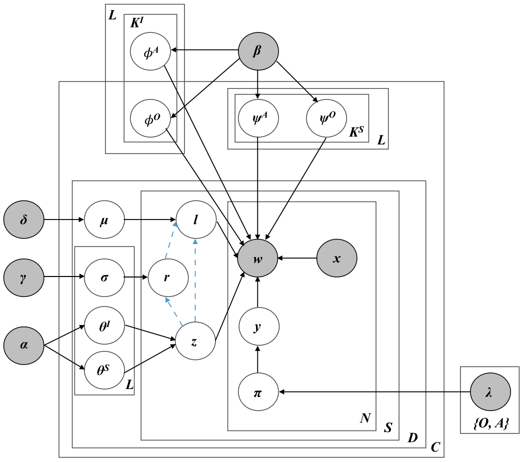


Fig. 4. Plate notation for AOS(exact).

The inference result for triplet (z, r, l) in the AOS(exact) model is as follows:

$$
\begin{aligned}
&P(z_{d,s} = k, r_{d,s} = j, l_{d,s} = t | \mathbf{z}_{\neg(d,s)}, \mathbf{r}_{\neg(d,s)}, \mathbf{l}_{\neg(d,s)}, \mathbf{y}, \mathbf{w}, \mathbf{x}) \\
&\quad \propto \frac{C^d_{(j)} + \gamma}{C^d_{(\cdot)} + 2\gamma} \times \frac{C^{d,j}_{(k)} + \alpha}{C^d_{(j)} + K^j\alpha} \times \frac{C^{d,j,k}_{(l)} + \delta}{C^{d,j}_{(k)} + L\delta} \\
&\quad \times \left(\frac{\Gamma\left(C^{A,t,j,k}_{(\cdot)} + V\beta\right)}{\Gamma\left(C^{A,t,j,k}_{(\cdot)} + N^{A,t,j,k}_{(\cdot)} + V\beta\right)} \cdot \prod_{v=1}^{V} \frac{\Gamma\left(C^{A,t,j,k}_{(v)} + N^{A,t,j,k}_{(v)} + \beta\right)}{\Gamma\left(C^{A,t,j,k}_{(v)} + \beta\right)} \right) \\
&\quad \times \left(\frac{\Gamma\left(C^{O,t,j,k}_{(\cdot)} + V\beta\right)}{\Gamma\left(C^{O,t,j,k}_{(\cdot)} + N^{O,t,j,k}_{(\cdot)} + V\beta\right)} \cdot \prod_{v=1}^{V} \frac{\Gamma\left(C^{O,t,j,k}_{(v)} + N^{O,t,j,k}_{(v)} + \beta\right)}{\Gamma\left(C^{O,t,j,k}_{(v)} + \beta\right)} \right) \qquad (3)
\end{aligned}
$$

where $C^d_{(j)}$ is the number of sentences assigned with type j (common/specific) topic in document d, $C^{d,j}_{(k)}$ is the number of sentences assigned to topic k of type j in document d, and $C^{d,j,k}_{(t)}$ is the number of sentences assigned to sentiment t of type j's topic k. The rest of math symbols are the same as those in 1.

References

[1] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang, "Cross-lingual mixture model for sentiment classification," in *Proc. ACL*, 2012, pp. 572–581.

[2] L. Gui *et al.*, "Cross-lingual opinion analysis via negative transfer detection," in *Proc. ACL*, Jun. 2014, pp. 860–865.

[3] G. Zhou, T. He, J. Zhao, and W. Wu, "A subspace learning framework for cross-lingual sentiment classification with partial parallel data," in *Proc. IJCAI*, 2015, pp. 1426–1433.

[4] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proc. ACL*, 2007, pp. 976–983.

[5] M. Long, J. Wang, G. Ding, W. Cheng, X. Zhang, and W. Wang, "Dual transfer learning," in *Proc. SDM*, 2012, pp. 540–551.

[6] F. Zhuang, P. Luo, C. Du, Q. He, and Z. Shi, "Triplex transfer learning: Exploiting both shared and distinct concepts for text classification," in *Proc. WSDM*, 2013, pp. 425–434.

[7] S. Li, S. Song, and G. Huang, "Prediction reweighting for domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1682–1695, Jul. 2017.

[8] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2006, pp. 120–128.

[9] Z. Lin, X. Jin, X. Xu, W. Wang, X. Cheng, and Y. Wang, "A cross-lingual joint aspect/sentiment model for sentiment analysis," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2014, pp. 1089–1098.

[10] Z. Lin *et al.*, "An unsupervised cross-lingual topic model framework for sentiment classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 432–444, Mar. 2016.

[11] X. Zhou, X. Wan, and J. Xiao, "Cross-lingual sentiment classification with bilingual document representation learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1403–1412.

[12] M. Paul and R. Girju, "Cross-cultural analysis of blogs and forums with mixed-collection topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2009, pp. 1408–1417.

[13] Y. Bao, N. Collier, and A. Datta, "A partially supervised cross-collection topic model for cross-domain text classification," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 239–248.

[14] F. Zhuang *et al.*, "Collaborative dual-PLSA: Mining distinction and commonality across multiple domains for text classification," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 359–368.

[15] L. Li, X. Jin, and M. Long, "Topic correlation analysis for cross-domain text classification," in *Proc. AAAI*, 2012.

[16] C. Ma, M. Wang, and X. Chen, "Topic and sentiment unification maximum entropy model for online review analysis," in *Proc. 24th Int. Conf. World Wide Web (WWW)*, 2015, pp. 998–1004.

[17] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proc. 4th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2011, pp. 815–824.

[18] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2009, pp. 375–384.

[19] S. Ben-david, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. NIPS*, 2006, pp. 137–144.

[20] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 210–219.

[21] T. Li, V. Sindhwani, C. Ding, and Y. Zhang, "Knowledge transformation for cross-domain sentiment classification," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2009, pp. 716–717.

[22] F. Zhuang, P. Luo, H. Xiong, Z. Shi, Q. He, and Y. Xiong, "Exploiting associations between word clusters and document classes for cross-domain text categorization," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2010, pp. 13–24.

[23] H. Wang, H. Huang, F. Nie, and C. Ding, "Cross-language Web page classification via dual knowledge transfer using nonnegative matrix tri-factorization," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. (SIGIR)*, 2011, pp. 933–942.

[24] F. Zhuang, P. Luo, P. Yin, Q. He, and Z. Shi, "Concept learning for cross-domain text classification: A general probabilistic framework," in *Proc. IJCAI*, 2013, pp. 1960–1966.

[25] X. Hu, J. Pan, P. Li, H. Li, W. He, and Y. Zhang, "Multi-bridge transfer learning," *Knowl.-Based Syst.*, vol. 97, pp. 60–74, Apr. 2016.

[26] S. Chandar *et al.*, "An autoencoder approach to learning bilingual word representations," in *Proc. NIPS*, 2014, pp. 1853–1861.

[27] H. Zhou, L. Chen, F. Shi, and D. Huang, "Learning bilingual sentiment word embeddings for cross-language sentiment classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 430–440.

[28] X. Zhou, X. Wan, and J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 247–256.

[29] A. M. Fernández, A. Esuli, and F. Sebastiani, "Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification," *J. Artif. Int. Res.*, vol. 55, no. 1, pp. 131–163, Jan. 2016.

[30] R. Xu and Y. Yang, "Cross-lingual distillation for text classification," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1415–1425.

[31] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: Are more languages better?" in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 28–36.

[32] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2008, pp. 553–561.

[33] X. Wan, "Co-training for cross-lingual sentiment classification," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. (AFNLP)*, 2009, pp. 235–243.

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[35] G. Doyle and C. Elkan, "Accounting for burstiness in topic models," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 281–288.

[36] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis, "A time-dependent topic model for multiple text streams," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 832–840.

[37] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, Apr. 2004.

[38] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 58–68.

[39] S. Jain and S. Batra, "Cross lingual sentiment analysis using modified BRAE," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 159–168.

[40] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.

[41] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid," in *Proc. EMNLP*, 2010, pp. 56–65.

[42] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed Gibbs sampling for latent Dirichlet allocation," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 569–577.

[43] P. Prettenhofer and B. Stein, "Cross-lingual adaptation using structural correspondence learning," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 1, pp. 1–22, Oct. 2011.

[44] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248547.1248548

[45] G. Heinrich, "Parameter estimation for text analysis," Fraunhofer IGD, Darmstadt, Germany, Tech. Rep., 2004.

Deqing Wang received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2013.

He is currently an Associate Professor with the School of Computer Science and the Deputy Chief Engineer with the National Engineering Research Center for Science Technology Resources Sharing and Service, Beihang University. His research focuses on text categorization and data mining for software engineering and machine learning.

Baoyu Jing received the bachelor's degree from Beihang University, Beijing, China, in 2016, and the master's degree from School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, in 2018.

His main research interests include topic model and cross-domain transfer learning.

Chenwei Lu received the master's degree in computer science from Beihang University, Beijing, China, in 2018.

His main research interests include transfer learning and sentiment analysis.

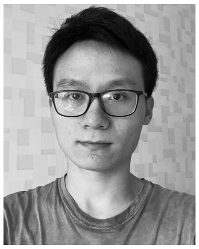
Junjie Wu is currently a Full Professor with the Information Systems Department, School of Economics and Management, Beihang University, Beijing, China. He is also the Director of the Research Center for Data Intelligence and the Vice Director of the Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations. His general area of research is data mining, with a special interest in social, urban, and financial computing.

Prof. Wu was a recipient of various national academic awards, including the NSFC Distinguished Young Scholars, the MOE Changjiang Young Scholars, and the National Excellent Doctoral Dissertation.

Guannan Liu received the Ph.D. degree from Tsinghua University, Beijing, China.

He is currently an Assistant Professor with the Department of Information Systems, Beihang University, Beijing. His research interests include data mining, business intelligence, and anomaly detection.

Dr. Liu's work has been published in journals, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Knowledge Discovery from Data, ACM Transactions on Intelligent Systems and Technology*, and *Decision Support Systems*, and also in conference proceedings, such as ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), International Conference on Data Mining (ICDM), and SIAM International Conference on Data Mining (SDM).



Chenguang Du is currently pursuing the master's degree in computer science with Beihang University, Beijing, China.

His main research interests include transfer learning and sentiment analysis.

Fuzhen Zhuang is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has published more than 70 articles in some prestigious refereed journals and conference proceedings, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CYBERNETICS, *ACM Transactions on Intelligent Systems and Technology*, *Information Sciences*, International Joint Conference on Artificial Intelligence (IJCAI), AAAI Conference on Artificial Intelligence (AAAI), International World Wide Web Conference (WWW), IEEE International Conference on Data Engineering (ICDE), ACM International Conference on Information and Knowledge Management (ACM CIKM), ACM International Conference on Web Search and Data Mining (ACM WSDM), SIAM International Conference on Data Mining (SIAM SDM), and IEEE International Conference on Data Mining. His research interests include transfer learning, multitask learning, and recommendation systems.