Modeling Language Discrepancy for Cross-Lingual Sentiment Analysis

Conference Paper · November 2017

DOI: 10.1145/3132847.3132915

Qiang Chen
AI Lab

Chenliang Li
Wuhan University

Modeling Language Discrepancy for Cross-Lingual Sentiment Analysis

Qiang Chen[1], Chenliang Li[2*], Wenjie Li[3], and Yanxiang He[2]
1. Tencent AI Lab, Shenzhen, 518000, China
ethanqchen@tencent.com
2. State Key Lab of Software Engineering, School of Computer, Wuhan University, Wuhan, 430072, China
{cllee, yxhe}@whu.edu.cn
3. Department of Computing, The Hong Kong Polytechnic University, 999077, Hong Kong
cswjli@comp.polyu.edu.hk

ABSTRACT

Language discrepancy is inherent and be part of human languages. Thereby, the same sentiment would be expressed in different patterns across different languages. Unfortunately, the language discrepancy is overlooked by existing works of cross-lingual sentiment analysis. How to accommodate the inherent language discrepancy in sentiment for better cross-lingual sentiment analysis is still an open question. In this paper, we aim to model the language discrepancy in sentiment expressions as intrinsic bilingual polarity correlations (IBPCs) for better cross-lingual sentiment analysis. Specifically, given a document of source language and its translated counterpart, we firstly devise a sentiment representation learning phase to extract monolingual sentiment representation for each document in this pair separately. Then, the two sentiment representations are transferred to be the points in a shared latent space, named hybrid sentiment space. The language discrepancy is then modeled as a fixed transfer vector under each particular polarity between the source and target languages in this hybrid sentiment space. Two relation-based bilingual sentiment transfer models (*i.e.,* RBST-s, RBST-hp) are proposed to learn the fixped transfer vectors. The sentiment of a target-language document is then determined based on the transfer vector between it and its translated counterpart in the hybrid sentiment space. Extensive experiments over a real-world benchmark dataset demonstrate the superiority of the proposed models against several state-of-the-art alternatives.

KEYWORDS

Language discrepancy; Bilingual sentiment transfer model; Cross-lingual sentiment analysis

1 INTRODUCTION

It is no doubt that the sentiment analysis in the low-resource languages requires massive human effort to construct the training data, which is expensive and tedious. Recently, cross-lingual sentiment analysis (CLSA) becomes a hot research topic in this field. The aim of CLSA is to leverage sentiment knowledge in a rich-resource (source) language to enhance sentiment analysis in the low-resource (target) language. For example, we can exploit a large number of training instances in English to improve the performance of sentiment analysis in Arabic or Chinese.

Many works have been proposed for CLSA tasks. Essentially, most of them address the task by studying the common patterns between sentiment expressions across languages. These works can be categorized into two main classes: traditional transfer learning (TTL) and cross-lingual representation learning (CLRL). The idea underlying TTL based methods is intuitive: TTL transfers the sentiment knowledge from the source language to the target language through the common sentiment features identified by using language translation techniques [7, 14, 20, 23]. CLRL based methods aim to infer a common latent feature space through the bilingual representation learning techniques [10, 25, 28, 29, 31]. The sentiment analysis is then conducted by using the latent representations. Both TTL and CLRL based methods require a bilingual parallel corpus of high quality. However, the construction of high-quality parallel corpus could take much human effort, which is expensive and unrealistic. Therefore, many works resort to adopting pseudo-parallel corpus built by using a machine translation service [7, 23].

The language discrepancy is inherent and be part of human languages. That is, sentiments are often expressed in different patterns across languages. Unfortunately, this language discrepancy is overlooked by existing CLSA works. Both TTL and CLRL based methods could suffer a lot from the problem of sentiment expression discrepancy, because different languages have different grammars and syntactic structures. Identifying the common sentiment features through language translation in TTL is not an easy task. It has been reported that a certain percentage of sentiments have been altered by machine translation services [4, 17]. On the other hand, due to the language discrepancy, different languages share some latent sentiment features but still have their own language-dependent features. The latent sentiment expression features

[*] Chenliang Li is the corresponding author.

CIKM'17 , November 6–10, 2017, Singapore, Singapore
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4918-5/17/11. . . $15.00
https://doi.org/10.1145/3132847.3132915

of languages S and T can be represented as $(\theta_0 + \theta_S)$ and $(\theta_0 + \theta_T)$ respectively, where θ_0 refers to the shared features, and θ_S and θ_T refer to the language-dependent features [18]. The CLRL based methods transform monolingual latent feature spaces (i.e., $(\theta_0 + \theta_S)$ and $(\theta_0 + \theta_T)$) to a shared bilingual latent feature space (i.e., θ_0) by compressing the language-dependent features (i.e., θ_S and θ_T) [10, 28, 31]. This methodology inevitably causes information loss, leading to unsatisfactory performance for sentiment analysis.

How to model this language discrepancy for better cross-lingual sentiment analysis is an interesting question. To this end, in this paper, we propose to model the language discrepancy in sentiment as the intrinsic bilingual polarity correlations (IBPCs) for CLSA. A **r**elation-based **b**ilingual **s**entiment **t**ransfer learning framework is proposed to learn the IBPC, named RBST. Figure 1 demonstrates the basic idea of RBST. In detail, we take pseudo-parallel document pairs as input. We first extract the high-level general semantics of each document by using a language specific CNN model in isolation (step ① in Figure 1). Then, the monolingual sentiment for each document is extracted from its general semantics through an orthogonal transformation method [21], also in monolingual mode (step ② in Figure 1). To model the language discrepancy in sentiment, we then map the resultant monolingual sentiments of documents of different languages into a shared latent space, named hybrid sentiment space. The hybrid sentiment space assumed here is totally different from the shared bilingual latent feature space assumed in CLRL based methods. Due to the fact that different languages express the same sentiment in different patterns, documents from different languages and polarities are automatically distributed to $g \times p$ areas in the hybrid sentiment space, where g refers to the number of languages and p refers to the number of polarities (i.e., $g = 2$, $p = 2$ is used in this paper). An area is denoted as $G(l)$, where G refers to language (*e.g.,* EN/CN[1]), and l refers to polarity (*e.g.,* +/-[2]). Here, we formulate the IBPCs as transfer vectors in the hybrid sentiment space. A transfer vector is defined as an offset vector that approximately transfers the representation of a document to the representation of its translated counterpart in the hybrid sentiment space (step 3 in Figure 1). In other words, a transfer vector for a particular polarity encodes the IBPC in the hybrid sentiment space.

Specifically, we propose two relation-based bilingual sentiment transfer learning processes (i.e., RBST-s and RBST-hp) to learn the two IBPCs respectively (i.e., EN(−)-to-CN(−) and EN(+)-to-CN(+)). While RBST-s directly learn the transfer vectors in the hybrid sentiment space, RBST-hp transfers sentiment representations further on the IBPC-specific hyperplanes where the learnt IBPC for each polarity can be more discriminative. Jointly learning the document sentiment representations and transfer vectors in the hybrid sentiment space enables the RBST models to avoid the aforementioned information loss problem ubiquitously exists in the CLRL based methods. Experimental results conducted on a benchmark dataset show that the proposed models outperform strong baselines and also several state-of-the-art methods. In summary, the main contributions of this paper are as follows:

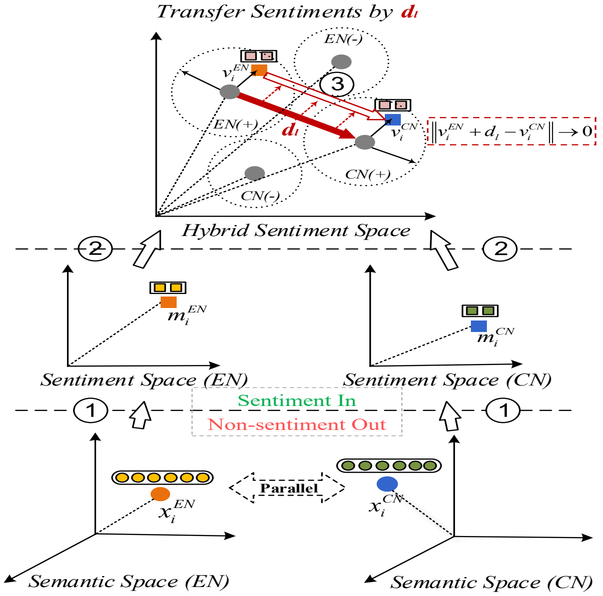


Figure 1: The basic idea of RBST framework. $\mathbf{d}_l$ refers to the IBPC between areas EN(+) and CN(+).

- To the best of our knowledge, we are the first to consider to model the language discrepancy in the task of CLSA. Specifically, the novelty lies in modeling the language discrepancy in CLSA as the intrinsic bilingual polarity correlations (IBPCs) in the latent Hybrid Sentiment Space.
- We present two relation-based bilingual sentiment transfer models, RBST-s and RBST-ph, to learn the IBPC as transfer vectors in the hybrid sentiment space. While RBST-s directly approximates the transfer vectors in the hybrid sentiment space, RBST-ph aims to learn the transfer vectors on IBPC-specific hyperplanes in the hybrid sentiment space, which leads to a more discriminative estimation of the transfer vectors.
- The extensive experiments over a real-world benchmark dataset validates the superiority of the proposed models. Supercharged by the modeling of IBPCs, we also observe that the monolingual sentiment representation learnt through RBST models carry more discriminative sentiment information than using the existing state-of-the-art bilingual representation learning techniques, leading to better sentiment classification performance.

[1] *EN* refers to English language, while *CN* refers to Chinese language.
[2] Symbol + refers to the positive sentiment, while − refers to the negative sentiment.

2 RELATED WORK

In the literature, there are two classes of methods for CLSA tasks: traditional transfer learning (TTL) and cross-lingual representation learning (CLRL).

TTL based methods transfer the sentiment knowledge from the source language to the target language through the common sentiment features identified by using language translation techniques. They include ensemble learning methods and standard transfer learning methods. Ensemble learning methods combine multiple classifiers trained over training data from both the source and target languages by voting mechanisms[22]. Standard transfer learning methods firstly build the language connection on a bilingual parallel or pseudo-parallel corpus. They then transfer the sentiment knowledge from the source language to the target language through the language connection built by using machine translation services or bilingual parallel corpus [14, 19, 20, 23]. TTL based methods would suffer from the language discrepancy problem, because they identify the common sentiment features through machine translation which inevitably alters sentiment [4, 17].

With the increased attention to distributed representation learning [11] and deep learning applications [9], CLRL based methods aim to infer a common latent feature space through the bilingual representation learning techniques. The sentiment analysis is then conducted by training extra classifiers using the latent representations. For example, both of Chandar A P et al. [1] and Zhou et al. [28] use two monolingual auto-encoders to learn the bilingual representations of bilingual parallel texts by sharing the hidden layers of the two auto-encoders, and similar methods are also proposed in [6, 12]. Later, Zhou et al. [30, 31] learn the bilingual sentimental representations of documents by incorporating the polar information of the labeled training data into learning. Also, several works employ non-negative matrix factorization models to infer the common latent sentiment features to generate the bilingual document representations [25, 26, 29]. However, the CLRL based methods inevitably causes information loss through the compression, leading to unsatisfactory performance for sentiment analysis.

The semantic relation can be completely modeled as the translation [15, 16]. However, it is not widely used in other NLP tasks except for entity relation learning in knowledge base. The semantic relation between two entities can be captured by translating from the one called the head to another called the tail in a semantic space (*i.e.,*, entity space or relation space). The relation is then specifically modeled as a translation vector. The prototype of translation based models is an unstructured model which treats semantic relations as zero vectors [2]. Later, Bordes et al. [3] define the semantic relation as a translation vector which is the subtraction from the vector representation of the tail entity to that of the head entity in the entity space, called the TransE model. To better capture 1-to-n relations, Wang et al. [24] propose to project entities into specific relation hyperplanes so that entities have distinct representations in different relation hyperplanes, which makes 1-to-n relation capturing possible, called the TranH model. In this paper, we are the first to introduce the translation based relation learning methods into CLSA task to model language discrepancy in sentiment as intrinsic bilingual polarity correlations (IBPCs).

The closest work is the DOC model proposed in [10]. It learns the common representation of bilingual parallel documents by approximating embeddings of semantically equivalent documents while maintaining sufficient distance between those of dissimilar texts. The learning is semantic-driven but not sentiment-driven. In comparison, the focus of our work is to learn the language discrepancy in sentiment using only limited labeled source-language documents. The discrepancy does not only directly help to determine the sentiment polarity of the document of the source language and its translated counterpart, but also guide the model to carry more discriminative sentiment information during document representation learning.

3 OUR APPROACH

As a preliminary study, we focus on bilingual-language sentiment analysis with two polarities[3]. We model the language discrepancy in sentiment as the **i**ntrinsic **b**ilingual **p**olarity **c**orrelations (IBPCs), and instantiate the latter as the transfer vectors in a shared latent feature space, named hybrid sentiment space. We first describe the proposed framework in detail, including the sentiment representation learning process, and two relation-based bilingual sentiment transfer models, RBST-s and RBST-ph (Section 3.1). Then, we describe the training scheme for our proposed models (Section 3.2). At last, we discuss sentiment predication by using RBST models (Section 3.2.1).

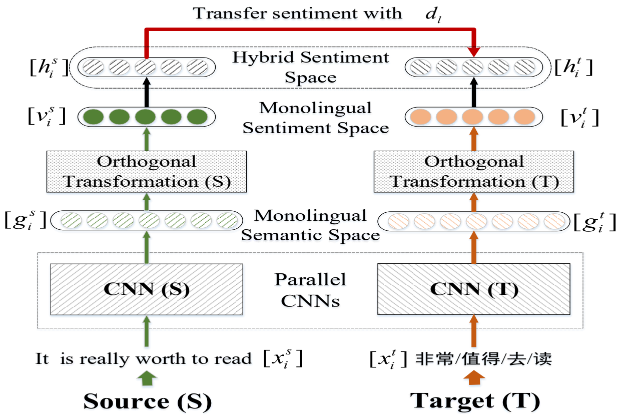
3.1 The Proposed Framework

We take pseudo-parallel bilingual corpus as the input to our proposed RBST models. Let the bilingual training corpus $X = \{(x_i^s, x_i^t)\}_{i=1}^M$ contain M training pseudo-parallel training pairs. x_i^t is the translation in target language t from document x_i^s of source language s, by using a machine translation service[4]. That is, x_i^t and x_i^s are considered to have the same sentiment label $\mathbf{y} \in \{-, +\}$. RBST models firstly learn the monolingual general semantic representations $(\mathbf{g}_i^s, \mathbf{g}_i^t)$ of (x_i^s, x_i^t) with two language-specific convolutional neural networks independently (*i.e.,* in monolingual mode). Then, we further adopt an orthogonal transformation method [21] to extract the latent sentiment features $(\mathbf{m}_i^s, \mathbf{m}_i^t)$ from $(\mathbf{g}_i^s, \mathbf{g}_i^t)$, also in monolingual mode. To measure the IBPC as a fixed transfer vector for the bilingual sentiment representations under each polarity, *i.e.,* $\mathbf{d}_l$ $(l \in \{-, +\})$, RBST models project $(\mathbf{m}_i^s, \mathbf{m}_i^t)$ to their positions $(\mathbf{v}_i^s, \mathbf{v}_i^t)$ in the hybrid sentiment space. The transfer vectors are then estimated by minimizing the average transfer loss over all bilingual parallel

[3] We use symbol +/- to refer to the positive and negative sentiment respectively.

[4] In this work, we use Google Translate at http://translate.google.com

training pairs. The overall framework of RBST models is illustrated in Figure 2.



Transfer sentiment with d_l
$[h_i^s]$
Hybrid Sentiment Space
$[h_i^t]$
$[v_i^s]$
Monolingual Sentiment Space
$[v_i^t]$
Orthogonal Transformation (S)
Orthogonal Transformation (T)
$[g_i^s]$
Monolingual Semantic Space
$[g_i^t]$
CNN (S)
Parallel CNNs
CNN (T)
It is really worth to read $[x_i^s]$
$[x_i^t]$ 非常/值得/去/读
Source (S)
Target (T)

Figure 2: The overall workflow of RBST framework.

3.1.1 Sentiment Representation Learning. Sentiment is one aspect of generic semantics [21]. To learn the monolingual sentiments $(\mathbf{m}_i^s, \mathbf{m}_i^t)$ of (x_i^s, x_i^t), we learn their general semantic representations $(\mathbf{g}_i^s, \mathbf{g}_i^t)$ with two language-specific convolutional neural networks. To simplify the description of the learning process, we *use $*$ as a placeholder for either the language s or the language t*. Hence, for a document $x_i^* = \mathbf{e}_{1:n}^* = \mathbf{e}_1^* \oplus \mathbf{e}_2^* \oplus ... \oplus \mathbf{e}_n^*$, where $\mathbf{e}_j^* \in \mathbb{R}^h$ is the pre-trained h-dimension word embedding of the j^{th} entry in x_i^*, and $\oplus$ refers to the concatenation operator. We use a convolution-pooling layer with F filters $W^* = \{\mathbf{w}_f^*\}_{f=1}^F$ to learn the general semantic $\mathbf{g}_i^*$ of x_i^*. The detailed calculation is as follows:

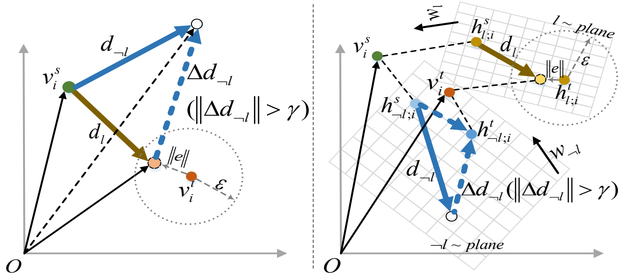
$$\begin{aligned}
&x_i^* = \mathbf{e}_{i,1:n}^* = \mathbf{e}_1^* \oplus \mathbf{e}_2^* \oplus ... \oplus \mathbf{e}_n^*;\\
&c_{f;i,j}^* = tanh(\mathbf{w}_f^* \mathbf{e}_{i,j+1:j+k}^* + b_f^*);\\
&\mathbf{c}_{f;i}^* = [c_{f;i,1}^*, ..., c_{f;i,n-k+1}^*];\\
&\mathbf{p}_{f;i}^* = [max\{\mathbf{c}_{f;i}^*\}, min\{\mathbf{c}_{f;i}^*\}, mean\{\mathbf{c}_{f;i}^*\}];\\
&\mathbf{g}_i^* = \mathbf{p}_{1;i}^* \oplus ... \oplus \mathbf{p}_{F;i}^*.
\end{aligned} \tag{1}$$

where k is the window of filter $\mathbf{w}_f^* \in \mathbb{R}^{1 \times k \cdot h}$. $\mathbf{c}_{f;i}^* \in \mathbb{R}^{n-k+1}$ refers to the new feature map of the convolution result under the filter $\mathbf{w}_f^*$. $\mathbf{p}_{f;i}^* \in \mathbb{R}^3$ represents the concatenation of the max-pooling, min-pooling and mean-pooling values over $\mathbf{c}_{f;i}^*$. Finally, $\mathbf{g}_i^* \in \mathbb{R}^{3F}$ is the general semantic representation of x_i^*, being concatenated by $\{\mathbf{p}_{f;i}^*\}_{i=1}^F$. Note that each (source/target) language has different parameters for the CNN model to extract general semantic representations.

As the intuition that sentiment space is the subspace of general semantic space, we then extract monolingual sentiment $\mathbf{m}_i^*$ from general semantic representation $\mathbf{g}_i^*$ using the orthogonal transformation method proposed in [21]. Specifically, sentiment information is extracted by using an orthogonal transformation as follows:

$$\mathbf{m}_i^* = \mathbf{Q}^* \mathbf{g}_i^* \tag{2}$$

where $\mathbf{Q}^* \in R^{3F \times 3F}$ is an orthogonal matrix.



(a) RBST-s

(b) RBST-ph

Figure 3: Illustrations of sentiment transfer in RBST-s and RBST-ph.

3.1.2 Relation-based Bilingual Sentiment Transfer Learning. Note that both $\mathbf{m}_i^s$ and $\mathbf{m}_i^t$ are calculated in the monolingual mode. In this sense, we can not measure the corresponding IBPC based on $\{\mathbf{m}_i^s, \mathbf{m}_i^t\}_{i=1}^M$ directly. Therefore, after the sentiment representation learning for the both two languages, we project bilingual parallel sentiments $(\mathbf{m}_i^s, \mathbf{m}_i^t)$ into a shared latent space, named hybrid sentiment space:

$$\mathbf{v}_i^s = \mathbf{P}^s \mathbf{m}_i^s \qquad \mathbf{v}_i^t = \mathbf{P}^t \mathbf{m}_i^t \tag{3}$$

where $(\mathbf{v}_i^s, \mathbf{v}_i^t)$ refers to the positions in the hybrid sentiment space for $(\mathbf{m}_i^s, \mathbf{m}_i^t)$, $\mathbf{P}^s$ and $\mathbf{P}^t$ are the transition matrices.

Due to the fact that different languages express the same sentiment in different patterns, it is expected that $\mathbf{v}_i^t$ and $\mathbf{v}_i^s$ should be distributed in their corresponding but different areas in the hybrid sentiment space. Hence, it is natural to model IBPC as the transfer vectors that bridge the mismatch between $\mathbf{v}_i^t$ and $\mathbf{v}_i^s$. Specifically, a transfer vector for a polarity is defined as an offset vector that approximates the difference between $\mathbf{v}_i^t$ and $\mathbf{v}_i^s$, given x_i^s contains the polarity. Here, we use a single transfer vector to accommodate the sentiment expression discrepancy for a polarity. In the following, we describe two relation-based bilingual sentiment transfer models, RBST-s and RBST-ph, to learn the transfer vectors. The modeling principles for RBST-s and RBST-ph are illustrated in Figure 3.

RBST-s Model. We estimate the transfer vectors from the parallel documents based directly on the positions of $\mathbf{v}_i^s$ and $\mathbf{v}_i^t$ in the hybrid sentiment space. Suppose the sentiment polarity of the parallel documents (x_i^s, x_i^t) is l. In ideal case, the sentiment representation $\mathbf{v}_i^t$ of x_i^t can be obtained by transferring the sentiment representation $\mathbf{v}_i^s$ of x_i^s with the transfer vector $\mathbf{d}_l$. That is, we have $\mathbf{v}_i^s + \mathbf{d}_l - \mathbf{v}_i^t = \mathbf{0}$. Thereby, we learn $\mathbf{d}_l$ by minimizing the transfer loss over all training parallel documents $\{(x_i^s, x_i^t)\}$ with polarity l. And the transfer loss is defined as follows:

$$f_l(\mathbf{v}_i^s, \mathbf{v}_i^t) = \|\mathbf{v}_i^s + \mathbf{d}_l - \mathbf{v}_i^t\|_2^2 \tag{4}$$

RBST-hp Model. We improve the RBST-s model by projecting all parallel documents (x_i^s, x_i^t) with polarity l into **p**olarity-specific **h**yperplanes, named RBST-ph. By further projecting sentiment information onto the polarity-specific hyperplane, we aim to find a suitable hyperplane in the hybrid sentiment space such that the IBPC between $\mathbf{v}_i^s$ and $\mathbf{v}_i^t$ is more discriminative. It ensures that the golden IBPC can be more clearly distinguished from the false one on the hyperplane. Given the normal vector w_l for the hyperplane H_l of polarity l, we calculate the shadow $\mathbf{h}_{l,i}^*$ of $\mathbf{v}_i^*$ as follows:

$$\mathbf{h}_{l,i}^* = \mathbf{v}_i^* - \mathbf{w}_l^\top \mathbf{v}_i^* \mathbf{w}_l \tag{5}$$

We learn $\mathbf{d}_l$ by minimizing the transfer loss over all training parallel documents $X^l = \{(x_i^s, x_i^t)\}$ with polarity l. And the transfer loss is defined by:

$$f_l(\mathbf{h}_i^s, \mathbf{h}_i^t) = \|\mathbf{h}_{l,i}^s + \mathbf{d}_l - \mathbf{h}_{l,i}^t\|_2^2 \tag{6}$$

3.2 Training Scheme

The RBST models are trained in a supervised manner. One straightforward approach is to specify the value of L_2 norm of the transfer vector $\mathbf{d}_l$ for each polarity l, and let the model learn all the parameters in a data-driven manner. However, this learning scheme mainly focuses on the discrepancy. Moreover, it is hard to find out the optimal L_2. These factors could lead model training to inferior local optimums.

Motivated by the noise contrastive estimation method [8], we adopt the margin-based ranking loss to learn the model parameters, by adding an auxiliary transfer vector $\mathbf{d}_{\neg l}$ for each polarity. In detail, the model learning process is supervised to find a local optimum such that $f_{\neg l}(\mathbf{v}_i^s, \mathbf{v}_i^t)$ is larger than $f_l(\mathbf{v}_i^s, \mathbf{v}_i^t)$ by a margin of γ, where $f_{\neg l}(\mathbf{v}_i^s, \mathbf{v}_i^t) = \|\mathbf{h}_i^s + \mathbf{d}_{\neg l} - \mathbf{h}_i^t\|_2^2$. The underlying idea is that the preferred optimum produces the smallest transfer loss, and any other candidate transfer vectors under the parameter setting will incur an average transfer loss larger than γ. According to Lemma 1, we can also show that the best $\mathbf{d}_{\neg l}$ has a minimal distance of γ to $\mathbf{d}_l$ in terms of L_2 norm. Hence, we believe that this constraint could enable us to learn the IBPC precisely in a data-driven manner.

Lemma 1. *If the golden polarity of (x_i^s, x_i^t) is l, then the contrastive transfer difference is $\Delta \boldsymbol{d}_{\neg l} = \boldsymbol{d}_{\neg l} - \boldsymbol{d}_l$, and $\|\Delta \boldsymbol{d}_{\neg l}\| \geq \gamma$, where γ is the margin.*

Proof 1. As shown in Figure 3(a): $\mathbf{v}_i^s + \mathbf{d}_l - \mathbf{v}_i^t \approx \mathbf{0}$, $\mathbf{d}_{\neg l} = \mathbf{d}_l + \Delta \mathbf{d}_{\neg l}$; Thus, $\mathbf{v}_i^s + \mathbf{d}_{\neg l} - \mathbf{v}_i^t \approx \Delta \mathbf{d}_{\neg l}$ and $\|\Delta \mathbf{d}_{\neg l}\| \geq \gamma$.

Thus, the objective function of RBST-s is a conditioned margin-based ranking loss, which is defined as:

$$\mathcal{L}_s = \sum_l \sum_{(x_i^s, x_i^t) \in X^l} \left[f_l(\mathbf{v}_i^s, \mathbf{v}_i^t) + \gamma - f_{\neg l}(\mathbf{v}_i^s, \mathbf{v}_i^t) \right]_+ + \beta \Theta \tag{7}$$
$$s.t. \quad \|\mathbf{v}_i^s\|_2 = 1, \qquad \|\mathbf{v}_i^t\|_2 = 1.$$

where X_l denotes the set of training parallel documents with polarity l, $[x]_+ \triangleq \max(0, x)$. $\gamma > 0$ is the margin hyperparameter. Θ is the sum of the L2 norms of model parameters, and β is the regularization coefficient. We handle the hard constraint $\|\mathbf{v}_i^s\|_2^2 = 1, \|\mathbf{v}_i^t\|_2^2 = 1$ by normalizing them after parameter updates in each learning epoch.

Similar to the learning of RBST-s, we introduce an auxiliary hyperplane $H_{\neg l}$ and an auxiliary transfer vector $d_{\neg l}$ for RBST-ph. The projection into the auxiliary hyperplane is calculated as follows:

$$\mathbf{h}_{\neg l,i}^* = \mathbf{v}_i^* - \mathbf{w}_{\neg l}^\top \mathbf{v}_i^* \mathbf{w}_{\neg l}; \tag{8}$$

where $\mathbf{w}_{\neg l}$ is the normal vector of $H_{\neg l}$, $\mathbf{h}_{\neg l,i}^*$ is the shadow of $\mathbf{v}_i^*$ on $H_{\neg l}$.

The objective function of RBST-ph is also a conditioned margin-based ranking loss function which is defined as:

$$\mathcal{L}_h = \sum_l \sum_{(x_i^s, x_i^t) \in X^l} \left[f_l(\mathbf{h}_{l,i}^s, \mathbf{h}_{l,i}^t) + \gamma - f_{\neg l}(\mathbf{h}_{\neg l,i}^s, \mathbf{h}_{\neg l,i}^t) \right]_+ + \beta \Theta$$
$$s.t. \begin{cases} \|\mathbf{h}_i^s\|_2 \leq 1, \quad \|\mathbf{h}_i^t\|_2 \leq 1; \\ \mid \mathbf{w}_l^\top \mathbf{d}_l \mid / \|\mathbf{d}_l\| \leq \sigma, \quad \mid \mathbf{w}_{\neg l}^\top \mathbf{d}_{\neg l} \mid / \|\mathbf{d}_{\neg l}\| \leq \sigma; \\ \|\mathbf{w}_l\| = 1, \quad \|\mathbf{w}_{\neg l}\| = 1. \end{cases} \tag{9}$$

where the former two constraints ensure the IBPC vectors $\mathbf{d}_l$ and $\mathbf{d}_{\neg l}$ being on the right hyperplanes (**Proof**. $\|\mathbf{h}_i^*\| = \|\mathbf{v}_i^*\| \cdot \cos(\theta) \leq \|\mathbf{v}_i^*\| = 1$, $|\mathbf{w}_l^\top \mathbf{d}_l| = 0 \leq \sigma$, where θ is the angle between $\mathbf{v}_i^*$ and $\mathbf{w}_l$). And the last two constraints guarantee the $\mathbf{w}_l$ and $\mathbf{w}_{\neg l}$ being normal vectors (**Proof**. if $\mathbf{d}_l$ in the hyperplane whose normal vector is $\mathbf{w}_l$, then $|\mathbf{w}_l^\top \mathbf{d}_l| = 0 \leq \sigma$). To optimize the objective function with hard constraints, we transform Equation (8) to its equivalent objective function with soft constraints.

$$\mathcal{L}_h = \sum_l \sum_{(x_i^s, x_i^t) \in X^l} \left[f_l(\mathbf{h}_{l,i}^s, \mathbf{h}_{l,i}^t) + \gamma - f_{\neg l}(\mathbf{h}_{\neg l,i}^s, \mathbf{h}_{\neg l,i}^t) \right]_+ + \beta \Theta$$
$$+ \lambda \left\{ \sum_{\upsilon \in \{\mathbf{h}_i^s, \mathbf{h}_i^t\}} \left[\|\upsilon\|_2^2 - 1 \right]_+ + \sum_{l \in \{-1, +1\}} \left[\frac{(\mathbf{w}_l^\top \mathbf{d}_l)^2}{\|\mathbf{d}_l\|_2^2} - \sigma^2 \right]_+ \right\} \tag{10}$$

where λ is the hyperparameter to control the importance of the soft constraints. $\left[\|\upsilon\|_2^2 - 1 \right]_+$ limits the range of sentiment representation.

The RBST-s and RBST-ph models are learned using mini-batch stochastic gradient descent (SGD) with the batch size 1. We randomly initialize each parameter matrix with samples following the uniform distribution $Unif(-\sqrt{\frac{6.}{(r+c)}}, \sqrt{\frac{6.}{(r+c)}})$, where r and c are the numbers of the rows and columns of the parameter matrix. Besides, we initialize the normal vectors of polarity-specific hyperplanes as unit vectors. Specifically, we initialize $\mathbf{Q}^{*(0)}$ with an orthogonal matrix. Nevertheless, $\mathbf{Q}^{*(t+1)}$ is not orthogonal after parameter update in each learning epoch. To ensure $\mathbf{Q}^*$ always being orthogonal, we reorthogonalize $\mathbf{Q}^{*(t+1)}$ as $\mathbf{U}^\top \mathbf{V}$ where $\mathbf{U}^\top \Sigma \mathbf{V}$ is the singular value decomposition result of $\mathbf{Q}^{*(t+1)}$, after each parameter updating like the work in [21] does. Meanwhile, sentiment vector representations are normalized in each learning epoch. The training process for RBST-ph model is illustrated in Algorithm 1. The training process for RBST-s is similar to RBST-ph by using $\mathcal{L}_s$ instead of $\mathcal{L}_h$.

3.2.1 Sentiment Prediction. For a bilingual parallel document pair (x_i^s, x_i^t), where x_i^s is the translation of x_i^t, we predict the polarity of x_i^t based on the difference between $(\mathbf{v}_i^s, \mathbf{v}_i^t)$ ($(\mathbf{h}_i^s, \mathbf{h}_i^t)$ in RBST-ph). Specifically, if $f_+(\mathbf{v}_i^s, \mathbf{v}_i^t) < f_-(\mathbf{v}_i^{\hat{s}}, \mathbf{v}_i^t)$ or $f_+(\mathbf{h}_i^{\hat{s}}, \mathbf{h}_i^t) < f_-(\mathbf{h}_i^{\hat{s}}, \mathbf{h}_i^t)$, then $l = +$; otherwise, $l = -$.

```
Algorithm 1: Our Approach
Input: Labeled bilingual parallel training data
    $X = (X^s, X^t)$, pre-trained embeddings $E^s$ and
    $E^t$, and hyperparameters $\{\alpha, \gamma, \lambda\}$
Output: Model parameters $\Theta =$
    $\{W^s, W^t, \mathbf{b}^s, \mathbf{b}^t, Q^s, Q^t, P^s, P^t, \mathbf{w}_l, \mathbf{w}_{\neg l}, \mathbf{d}_{-1}, \mathbf{d}_{+1}\}$
for $t = 1 : T$ do
    foreach $X_b \subseteq X$ do
        /* train model over all mini-batch $X_b$ */
        initialize objective function $\mathcal{L}_h = 0$;
        foreach $(x_i^s, x_i^t) \in X_b$ do
            $(\mathbf{g}_i^s, \mathbf{g}_i^t) \leftarrow CNNs(x_i^s, x_i^t)$;   /* Formula (1) */
            $(\mathbf{m}_i^s, \mathbf{m}_i^t) \leftarrow OT(\mathbf{g}_i^s, \mathbf{g}_i^t)$;   /* Equation 2 */
            /* hybrid sentiment space projection */
            $(\mathbf{v}_i^s, \mathbf{v}_i^t) \leftarrow Map(\mathbf{m}_i^s, \mathbf{m}_i^t)$;
            /* hyperplane projection */
            $(\mathbf{h}_i^s, \mathbf{h}_i^t) \leftarrow Transform((\mathbf{v}_i^s, \mathbf{v}_i^t))$;
            calculate $f_l(\mathbf{h}_i^s, \mathbf{h}_i^t)$ and $f_{\neg l}(\mathbf{h}_i^s, \mathbf{h}_i^t)$;
            /* Equation (6) */
            /* calculate noise-contrastive loss */
            calculate
            $\mathcal{L}_{h,i} \leftarrow [f_l(\mathbf{h}_i^s, \mathbf{h}_i^t) + \gamma - f_{\neg l}(\mathbf{h}_i^s, \mathbf{h}_i^t)]_+$;
            /* update objective function */
            $\mathcal{L}_h \leftarrow \mathcal{L}_h + \mathcal{L}_{h,i}$;
        $\Theta^{(t+1)} = \Theta^{(t)} - \alpha \cdot \frac{\partial \mathcal{L}_h}{\partial \Theta}$; /* update parameters */
        /* reorthogonalize $Q$, where
        $U^\top \Sigma V = Q^{*(t+1)}$ */
        $\overline{Q}^{*(t+1)} = U^\top V \ (* \in \{s, t\})$;
Return the optimal model parameters $\Theta^\diamond$.
```

For fine-grained sentiment analysis (*i.e.,* more than two polarities), we choose the polarity label that minimizes the transfer loss as follows:

$$l^\diamond = \arg\max_l f_l(\mathbf{v}_i^s, \mathbf{v}_i^t)$$
$$or$$
$$l^\diamond = \arg\max_l f_l(\mathbf{h}_{l,i}^s, \mathbf{h}_{l,i}^t)$$

Recall in Section 3.1.1 that sentiment representations $(\mathbf{m}_i^s, \mathbf{m}_i^t)$ are extracted independently. However, these two parallel procedures are coupled in relation-based bilingual sentiment transfer learning processes (*i.e.,* RBST-s and RBST-ph). This joint learning scheme transfers the supervision of IBPC modeling into both sentiment representation learning procedures. If there are no translations $\{x_i^s\}_{i=1}^M$ available for $\{x_i^t\}_{i=1}^M$ in hand, we can use RBST-s or RBST-ph to obtain the sentiment representations for target language data to train a sentiment classifier. In our work, the experimental results demonstrate that using sentiment representation $\mathbf{m}_i^t$ by RBST models delivers better sentiment classification performance than the existing bilingual representation learning based alternatives and TTL based alternatives (see Section 4.2).

4 EXPERIMENTS AND ANALYSIS

4.1 Dataset and Experimental Setting

Dataset. We evaluate the proposed models on an open CLSA benchmark dataset[5] [4, 7, 30, 32] in which the source language is English and the target language is Chinese. The provided dataset is a collection of bilingual Amazon product reviews in Books, DVD and Music domains. It contains 4,000 labeled English reviews for training, 4,000 Chinese reviews for testing, and 17,814, 47,071, 29,677 unlabeled Chinese reviews in three different domains respectively. The training and testing data are translated to the counterpart language respectively by using Google Translator. For the experiment comparison in each domain, we randomly select 500 reviews from test data as the development set. Moreover, for each domain, we put the unlabeled Chinese reviews and 5 millions Chinese posts crawled from Weibo[6] together to pre-train Chinese word embeddings using the Word2Vec tool[7]. We also leverage extra English Amazon product reviews provided in [13] to pre-train English word embeddings[8]. The dimension of word embeddings is 300.

Parameter Setting. For both the RBST-s and RBST-hp models, we empirically set the regularization coefficient $\beta = 10^{-6}$, the dimension of sentiment representation $z = 50$, the filter number $F = 100$ and the filter window $k = 3$. We use the grid search algorithm to choose the learning rate α from $\{0.001, 0.009, 0.01, 0.1\}$, the margin γ from $\{0.1, 1, 2, 3, 4, 5\}$ and the soft constraint hyperparameter λ from $\{0, 0.1, 0.25, 0.5, 1.0\}$. According to the experimental results on development sets, the optimal configurations are set as follows: $\alpha = 0.009$, $\gamma = 3$, $\lambda = 1.0$ in the BOOKs domain; $\alpha = 0.01$, $\gamma = 3$, $\lambda = 0.25$ in the DVD domain; $\alpha = 0.009$, $\gamma = 3$, $\lambda = 0.25$ in the MUSIC domain.

Methods in Comparison. We compare the proposed RBST models against the following recent state-of-the-art CLSA alternatives.

MT-SVM: It is a basic label-based transfer learning method. It uses Chinese translations of labeled English training data to train a Chinese SVM classifier.

CoTrain [23]: It is a standard co-training framework. We implement it using the same settings used in [23].

BiDOC [10]: It learns the compositional semantic representation of bilingual documents by pushing their representations to be close to each other. We implement this method

[5] The task 3 in NLP&CC 2013, and the dataset can be downloaded at: http://tcci.ccf.org.cn/conference/2013/dldoc/evdata03.zip.
[6] http://www.weibo.com
[7] Available at http://radimrehurek.com/gensim/
[8] The pre-trained word embedding models can be downloaded at: https://drive.google.com/open?id=0B0l0oLL2GUuoblNta0QyY1BkdGM.

Domain	TTL Framework		CLRL Framework				RBST Framework			
	MT-SVM	**CoTrain**	**BiDOC**	**BSWE**	**BiAtt**	**BiCNN**	**RBST-s**	**RBST-CN**	**RBST-Bi**	**RBST-hp**
<u>BOOKs</u>	0.735	0.798	0.783	0.811	0.821	0.801	0.807	0.806	0.812	0.819
<u>DVD</u>	0.760	0.775	0.795	0.816	0.837	0.816	0.824	0.825	0.829	0.835
<u>MUSIC</u>	0.739	0.772	0.771	0.794	0.813	0.789	0.799	0.794	0.812	0.814
mAcc	0.744	0.782	0.783	0.807	**0.824**	0.802	0.810	0.808	0.817	<u>0.823</u>
average	0.763		<u>0.804</u>				**0.815**			

Table 1: Macro performances in three domains. The bold font refers to the best performance while the underline refers to the second best. The macro performance of BiAtt model is reported in [32].

by replacing the nonlinear document composition function with a convolution-pooling layer pair for better performance.

BSWE [30]: This method learns the bilingual sentiment word embedding by integrating the polarity information in the training data.

BiAtt [32]: It uses bidirectional LSTM to learn the monolingual representations of the documents in both source language and target language with hierarchical attention layers for sentiment extraction.

BiCNN: It is a variation of RBST models. It does not project the monolingual representations into the hybrid sentiment space for IBPC learning, but directly concatenates them as the bilingual representation of bilingual parallel documents, like **Bi-Att** does.

RBST-CN: It trains the RBST-hp model, and then uses the resultant sentiment representations of the Chinese documents to train a SVM classifier.

RBST-Bi: It trains the RBST-hp model, and then uses the concatenation of the resultant sentiment representations of bilingual parallel documents to train a SVM classifier.

The comparison methods are categorized into three classes: the first two methods are strong TTL based methods; the next four methods are recent state-of-the-art cross-lingual representation learning based methods; and the last two methods are representation learning based solutions varied from RBST-ph model. For methods which train extra classifiers, we use support vector machines (SVMs) as basic classifiers. We use the Liblinear package [5] with linear kernel[9]. For MT-SVM and CoTrain, we use unigram+bigram feature setting as input. For all baseline methods, the parameters are primarily set as those used in their corresponding papers, and then are slightly tuned for better performance.

Evaluation Protocol. Following the existing works [4, 32], we adopt three metrics in terms of accuracy for performance evaluation.

$$\begin{aligned} Acc(f) &= \frac{n^f}{N^f}, \\ mAcc_m(F) &= \frac{1}{3} \cdot \sum_{f' \in \mathcal{F}} Acc(f'), \\ average(M) &= \frac{1}{\|M\|} \cdot \sum_{m \in M} mAcc(F) \end{aligned} \tag{11}$$

where n^f is the number of correct predictions and N^f is the total number of the test data; $f \in \{Books, DVD, Music\}$ is the domain set. $mAcc_m(F)$ is the mean accuracy of the method m over all domains, and $average(M)$ refers to the average accuracy over all methods for each framework.

4.2 Experimental Results and Analysis

Tables 1 and 2 show the macro- and micro-performances of all methods respectively. The first two methods are strong TTL based baselines which build language connections between bag-of-word features across languages using machine translation services. CoTrain performs much better than MT-SVM due to its iterative boosting learning. The middle four methods are state-of-the-art CLRL based methods which learn the bilingual representations from bilingual parallel documents. BiAtt outperforms the other three CLRL based methods by using a complex neural network architecture with a hierarchical sentiment attention mechanism. It is obvious that CLRL and RBST based methods outperform TTL based methods by a large margin, which suggests the superiority of deep learning techniques in sentiment learning. The last four RBST models (i.e., RBST-s, RBST-CN, RBST-Bi, RBST-hp) get better performances than most of the comparison methods. This demonstrates that the proposed sentiment expression discrepancy modeling in RBST models is effective for CLSA task. We further conduct detailed analysis based on the experimental results in the following.

Specifically, BiDOC, BSWE, RBST-CN and RBST-Bi all use simple deep learning architectures to derive the sentiment representation for a parallel documents. It is expected that BiDOC performs the worst, because it does not incorporate sentiment information into representation learning. Although, both BSWE and RBST-Bi incorporate sentiment information into representation learning process and use the bilingual representations to train a SVM classifier for prediction, RBST-Bi performs better than BSWE by an absolute increase of 1.0%. The bilingual sentiment representation of a document in BSWE is through a linear composition of bilingual sentiment word embeddings. This simple composition strategy can not fully express the sentiment information of a document due to the nonlinear structure of document. Besides, RBST-Bi learns the sentiment representation through a joint learning scheme, guided by the IBPC in the hybrid sentiment space. Note that both BiDOC and BSWE forcibly push bilingual parallel documents to one single point in a shared bilingual space by compressing the latent language-dependent features. They inevitably cause the information loss problem. Compared with RBST-Bi, their inferior performance demonstrates that

[9] The parameter setting used in this paper is *'-s 7'* which is same with the setting used in [4].

BOOKs	Positive			Negative			Acc
	P	R	F1	P	R	F1	
MT-SVM	0.778	0.665	0.715	0.706	0.805	0.752	0.735
CoTrain	0.803	0.790	0.796	0.793	0.806	0.800	0.798
BiDOC	0.771	0.806	0.788	0.796	0.761	0.778	0.783
BSWE	0.821	0.794	0.807	0.800	0.828	0.814	0.811
BiAtt	—	—	—	—	—	—	**0.821**
BiCNN	0.782	**0.844**	0.811	**0.830**	0.765	0.796	0.804
RBST-s	0.797	0.824	0.810	0.817	0.790	0.804	0.807
RBST-CN	0.845	0.744	0.793	0.772	**0.868**	0.817	0.806
RBST-Bi	**0.848**	0.766	0.803	0.786	0.857	0.820	0.812
RBST-hp	0.827	0.805	**0.816**	0.810	0.832	**0.821**	0.819

DVD	Positive			Negative			Acc
	P	R	F1	P	R	F1	
MT-SVM	0.769	0.744	0.756	0.752	0.777	0.764	0.760
CoTrain	0.776	0.774	0.775	0.775	0.776	0.775	0.775
BiDOC	0.843	0.725	0.779	0.759	0.865	0.808	0.795
BSWE	0.837	0.797	0.817	0.806	0.835	0.825	0.816
BiAtt	—	—	—	—	—	—	**0.837**
BiCNN	0.823	**0.817**	0.820	0.818	0.825	0.821	0.821
RBST-s	0.843	0.797	0.819	0.807	0.852	0.829	0.824
RBST-CN	0.842	0.800	0.821	0.810	0.850	0.829	0.825
RBST-Bi	0.848	0.802	0.824	0.812	0.856	0.833	0.829
RBST-hp	**0.852**	0.810	**0.830**	**0.819**	**0.859**	**0.839**	0.835

MUSIC	Positive			Negative			Acc
	P	R	F1	P	R	F1	
MT-SVM	0.790	0.651	0.714	0.703	0.827	0.760	0.739
CoTrain	0.789	0.744	0.766	0.758	0.801	0.779	0.772
MT-D2V	0.747	0.820	0.782	0.800	0.723	0.759	0.771
BSWE	**0.818**	0.756	0.786	0.773	**0.832**	0.802	0.794
BiAtt	—	—	—	—	—	—	0.813
BiCNN	0.775	**0.836**	0.804	**0.822**	0.757	0.788	0.796
RBST-s	0.814	0.774	0.794	0.785	0.824	0.804	0.799
RBST-CN	0.789	0.805	0.796	0.800	0.783	0.792	0.794
RBST-Bi	0.812	0.811	0.811	0.811	0.813	**0.812**	0.812
RBST-hp	0.808	0.824	**0.816**	0.820	0.804	**0.812**	**0.814**

Table 2: Micro performance. P: Precision, R: Recall, F1: micro-F measure, Acc: Accuracy. The performance of BiAtt model is directly cited, but [32] does not report the micro performance. - represents the unavailable values. The underline refers to the second best.

information loss is indeed an obstacle underlying the existing CLRL based methods. In addition, RBST-CN only uses the sentiment representations of Chinese documents to train an SVM classifier. Its superior performance over BiDOC and BSWE suggests that the RBST models can better uncover the underlying sentiment information to its fullness.

BiAtt is a deep learning method using a bi-directional LSTM model with hierarchical sentiment attention layers for sentiment representation learning in both the source language and the target language. The bidirectional memory neural network with hierarchical attention mechanism has being proved to be greatly superior to other models in document representation learning [27]. In comparison, RBST-hp learns the document representation using simple CNNs with only one convolution-pooling layer pair, but it studies the language discrepancy in sentiment to help analyzing the cross-lingual sentiment polarity. As shown in Tables 1 and 2, RBST-hp achieves the comparable performance to that of BiAtt in all domains. It demonstrates the disadvantage of RBST-hp in sentiment representation learning is filled up by the IBPC learning for cross-lingual sentiment analysis.

We also make another interesting observation. Although RBST-Bi and RBST-ph have the same training process, the corresponding prediction schemes are completely different. RBST-Bi trains an SVM-based sentiment classifier using the sentiment representations $(\mathbf{m}_i^s, \mathbf{m}_i^t)$. On the contrary, RBST-ph predicts the sentiment polarity according to the IBPC of the parallel documents of interest. As shown in Tables 1 and 2, RBST-ph performs better than RBST-Bi. Recall in Section 3.2, the objective function of RBST-ph is not designed to learn a decision boundary that distinguishes the positive and negative training instances to its maximum. That is, the underlying philosophy of RBST model is totally different to other discriminative models compared here (*e.g.,* SVM classifiers). The superiority of RBST-ph suggests that modeling language discrepancy in sentiment expression is beneficial to CLSA task, and the learning strategies proposed in RBST models are effective to harness this language discrepancy.

Finally, as shown in Tables 1 and 2, RBST-ph performs better than RBST-s. Recall that RBST-ph further transforms sentiment information from the hybrid sentiment space to polarity-specific hyperplanes where the estimated IBPC is more discriminative (see Section 3.1.2). The positive performance gain by using polarity-specific hyperplane validates that more discriminative IBPC is learnt by RBST-ph model.

4.3 Existence of IBPCs

Figure 4 plots the distributions of the training instances in the hybrid sentiment spaces in three domains, where we use singular value decomposition technique over the sentiment representations in the hybrid sentiment space for the visualization [25]. In each domain, the training instances are distributed in four areas, EN(+), CN(+), EN(−), CN(−). This observation is consistent with our 4-area hypothesis for the hybrid sentiment space. Additionally, we also mark the two transfer vectors between the bilingual areas for each polarity, *i.e.,* (EN(+), CN(+)) or (EN(−), CN(−)). This result confirms the existence of IBPC as the transfer vectors in the hybrid sentiment space. Another interesting observation is that the two transfer vectors in each domain are nearly orthogonal, which is consistent to the oppositeness between the negative and positive sentiments. This suggests that the proposed RBST models can successfully extract the sentiment information from the parallel documents.

4.4 Parameter Study

We further study the performance of RBST models by varying λ and γ. When studying a particular parameter, we fix the other parameter to the values used in Section 4.2.

4.4.1 Impact of Soft Constraints λ. The importance of soft constraints in RBST-ph learning is controlled by parameter λ. Figure 5(a) shows the curves of the performance fluctuations of RBST-hp by varying λ from 0.01 to 2.0 in all domains. The optimal performance is achieved when $\lambda = 1.0$ in the Books domain and $\lambda = 0.25$ in the DVD and Music domains. We observe that given weak constraints in the range of the representations of bilingual parallel sentiments (*i.e.,* $\lambda \leq 0.1$), the performance of RBST-ph degrades significantly.

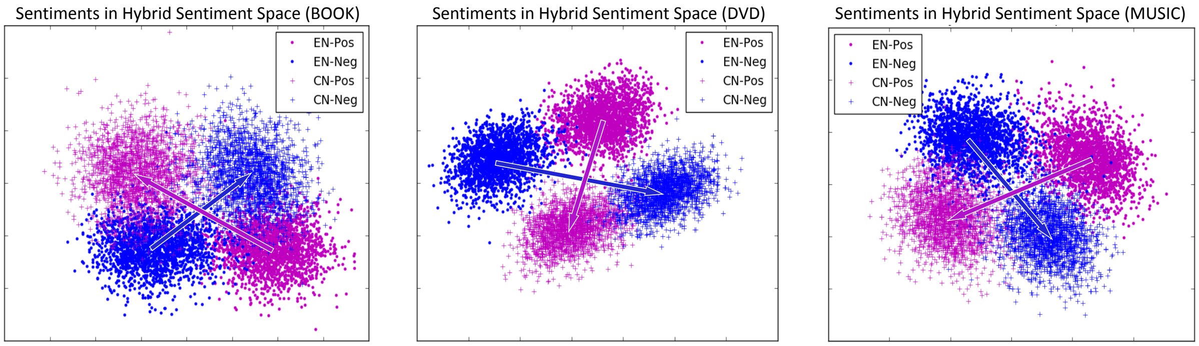
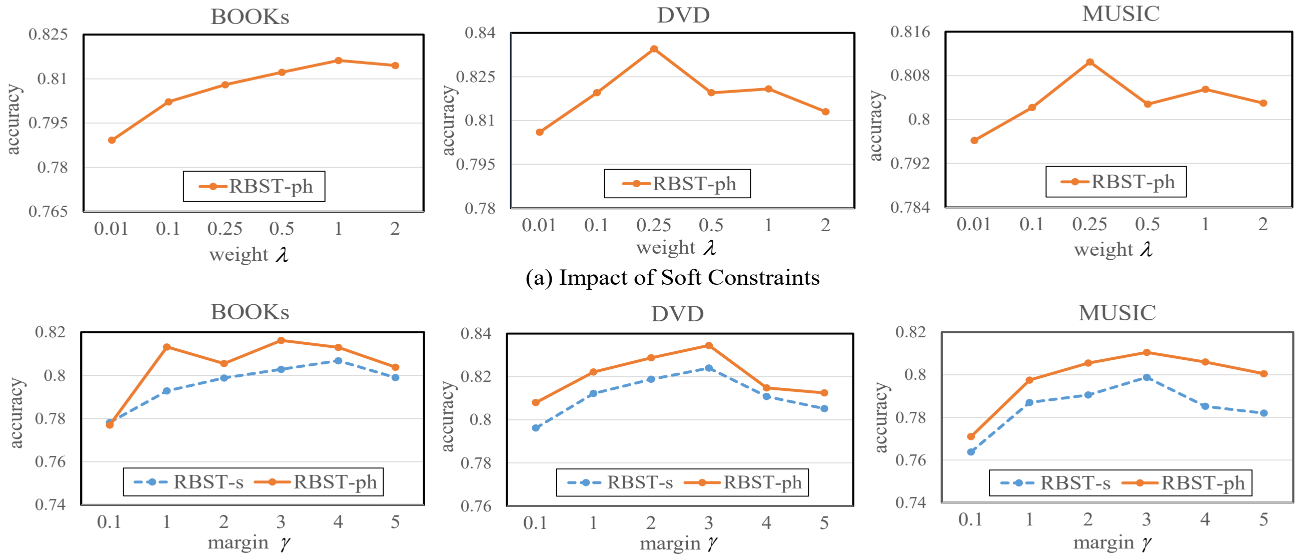


Figure 4: Distribution of training instances in the hybrid sentiment space for DVD domain. The arrows refer to the intrinsic polarity expression discrepancies across languages.



(a) Impact of Soft Constraints

(b) Impact of Margin for Error Tolerance

Figure 5: Impact of soft constraints λ (a) and margin γ (b).

This is reasonable since the model learning has a relatively larger search space, which takes more time to converge. Also, the resultant larger search space makes RBST-ph learning process to be more likely to reach an inferior local optimum. Besides, the classification performance also deteriorates a lot when λ becomes relative much larger (*i.e.,* $\lambda > 1.0$ in the Book domain and $\lambda > 0.25$ in the DVD and Music domains). Recall in Equation 9, a larger λ severely restrict the range of the projected representations in the polarity-specific hyperplanes. A stricter constraint could not absorb the possible noise information, leading to significant performance degradation. Based on the results, we set $\lambda = 1.0$ for the Book domain, and $\lambda = 0.25$ for the DVD and Music domains in our experiments.

4.4.2 Impact of Margin γ. Figure 5(b) plots the performance patterns by varying margin γ from 0.1 to 5. The optimal performance is achieved when margin $\gamma = 3$ in all three domains. The performances of the both RBST models in all domains decrease significantly when $\gamma \leq 1$. When γ is small, many candidate transfer vectors can match the objective function. In this sense, the learning process is easy to find a local but inferior optimum. Also, a larger γ could hinder the IBPC learning process such that the optimal transfer vectors are missed, because the contrastive loss difference is not matched with the margin. In detail, margin γ should satisfy the condition: $\gamma \leq 4$. The theoretical proof for RBST-s is as follows: since $\|\Delta \mathbf{d}_{\neg l}\| = \|\mathbf{v}_i^s + \mathbf{d}_{\neg l} - \mathbf{v}_i^t\| \leq \|\mathbf{v}_i^s\| + \|\mathbf{d}_{\neg l}\| + \|\mathbf{v}_i^t\|$, and $\|\mathbf{d}_{\neg l}\| = \|\mathbf{v}_i^t - \mathbf{v}_i^s\| \leq \|\mathbf{v}_i^t\| + \|\mathbf{v}_i^s\| = 2$, then we can get $\|\Delta \mathbf{d}_{\neg l}\| \leq 4$. Based on $\|\Delta \mathbf{d}_{\neg l}\| \geq \gamma$, we obtain $\gamma \leq 4$. The same condition can also be met in a similar derivation for RBST-ph model.

5 CONCLUSION AND FUTURE WORK

In this paper, we model the language discrepancy in sentiment as intrinsic bilingual sentiment correlations (IBPC) for better bilingual sentiment analysis. Based on this, we propose a novel relation-based bilingual sentiment transfer learning framework to solve CLSA problems. Specifically, we propose two RBST based models (*i.e.,* RBST-s and RBST-ph) to learn the IBPC as the transfer vectors in the hybrid sentiment space (*i.e.,* EN(−)-to-CN(−) and EN(+)-to-CN(+)). We then use the learned IBPC to directly determine the sentiment polarity of a target-language document with its source-language translation. The experiments over a real-world benchmark dataset suggest that the proposed IBPC framework is effective and robust for the task of CLSA. The joint learning process guided by the IBPC learning also deliver a better representation learning for sentiment information.

The sentiment extraction procedure in the two propsoed RBST models is based on a simple CNN architecture followed by an orthogonal transformation. We plan to extend the RBST models to exploit the word sequence modeling offered by RNN models, *e.g.,* a LSTM + attention architecture. The IBPC examined in this paper only includes two relations: EN(−)-to-CN(−) and EN(+)-to-CN(+). As a future work, we plan to adapt RBST models to address the task of multilingual sentiment analysis with fine-grained polarities.

ACKNOWLEDGEMENT

We thank all the anonymous reviewers for their detailed and insightful comments on this paper. This research was supported by National Natural Science Foundation of China (No. 61502344, No. 61672445, No. 61472290), Natural Scientific Research Program of Wuhan University (No. 2042017k-f0225, No. 2042016kf0190), Academic Team Building Plan for Young Scholars from Wuhan University (No. Whu2016012).

REFERENCES

[1] Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of NIPS 2014*. pages 1853–1861.

[2] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In *Proceedings of International Conference on Artificial Intelligence and Statistics 2012*. pages 127–135.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of NIPS 2013*. pages 2787–2795.

[4] Qiang Chen, Wenjie Li, Yu Lei, Xule Liu, and Yanxiang He. 2015. Learning to adapt credible knowledge in cross-lingual sentiment analysis. In *Proceedings of ACL 2015*. pages 419–429.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *JMLR:* volume 9 (2008), pages 1871–1874.

[6] Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML 2015*. pages 748–756.

[7] Lin Gui, Ruifeng Xu, Qin Lu, Jun Xu, Jian Xu, Bin Liu, and Xiaolong Wang. 2014. Cross-lingual Opinion Analysis via Negative Transfer Detection. In *Proceedings of ACL 2014*. pages 860–865.

[8] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.. In *Proceedings of International Conference on Artificial Intelligence and Statistics 2010*. pages 297–304.

[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of WWW 2017*. pages 173–182.

[10] Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of ACL 2014*. pages 58–68.

[11] Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceeding of ICML 2014*. pages 1188–1196.

[12] Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing 2015*. pages 151–159.

[13] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of SIGIR 2015*. pages 43–52.

[14] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of ACL 2012*. pages 572–581.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*. pages 3111–3119.

[16] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations.. In *Proceedings of NAACL 2013*. pages 746–751.

[17] Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2015. How translation alters sentiment. *JAIR:* volume 1 (2015), pages 1–20.

[18] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *TKDE:* volume 22, 10 (2010), pages 1345–1359.

[19] Kashyap Popat and Balamurali AR. 2013. The haves and the have-nots: Leveraging unlabelled corpora for sentiment analysis. In *Proceedings of ACL 2013*. pages 412–422.

[20] Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of ACL 2010*. pages 1118–1127.

[21] Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense Word Embeddings by Orthogonal Transformation. In *Proceedings of NAACL 2016*. pages 767–777.

[22] Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of EMNLP 2008*. pages 553–561.

[23] Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL 2009*. pages 235–243.

[24] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of AAAI 2014*. pages 1112–1119.

[25] Min Xiao and Yuhong Guo. 2013. A novel two-step method for cross language representation learning. In *Proceedings of NIPS 2013*. pages 1259–1267.

[26] Min Xiao and Yuhong Guo. 2014. Semi-supervised matrix completion for cross-lingual text classification. In *Proceedings of AAAI 2014*. pages 1607–1613.

[27] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of ACL 2016*. pages 1480–1489.

[28] Guangyou Zhou, Tingting He, and Jun Zhao. 2014. Bridging the Language Gap: Learning Distributed Semantics for Cross-Lingual Sentiment Classification. In *Proceedings of NLP&CC 2014*. pages 138–149.

[29] Guangyou Zhou, Tingting He, Jun Zhao, and Wensheng Wu. 2015. A subspace learning framework for cross-lingual sentiment classification with partial parallel data. In *Proceedings of IJCAI 2015*. pages 1426–1432.

[30] Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of ACL 2015*. pages 430–440.

[31] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning. In *Proceedings of ACL 2016*. pages 1403–1412.

[32] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Attention-based LSTM Network for Cross-Lingual Sentiment Classification. In *Proceedings of EMNLP 2016*. pages 247–256.