Knowledge-Based Systems 107 (2016) 129–141

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Cross-lingual sentiment classification: Similarity discovery plus training data adjustment

Peng Zhang[a], Suge Wang[a,b,*], Deyu Li[a,b]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006, China

ARTICLE INFO

Article history:
Received 4 February 2016
Revised 13 May 2016
Accepted 5 June 2016
Available online 8 June 2016

Keywords:
Topic model
Cross-lingual sentiment classification
Semi-supervised learning

ABSTRACT

The performance of cross-lingual sentiment classification is sharply limited by the language gap, which means that each language has its own ways to express sentiments. Many methods have been designed to transmit sentiment information across languages by making use of machine translation, parallel corpora, auxiliary unlabeled samples and other resources. In this paper, a new approach is proposed based on the selection of training data, where labeled samples highly similar to the target language are put into the training set. The refined training samples are used to build up an effective cross-lingual sentiment classifier focusing on the target language. The proposed approach contains two major strategies: the aligned-translation topic model and the semi-supervised training data adjustment. The aligned-translation topic model provides a cross-language representation space in which the semi-supervised training data adjustment procedure attempts to select effective training samples to eliminate the negative influence of the semantic distribution differences between the original and target languages. The experiments show that the proposed approach is feasible for cross-language sentiment classification tasks and provides insight into the semantic relationship between two different languages.

1. Introduction

As Internet access has become globally convenient, our world has experienced the development of a new fashion of social media, in which the volume of user-generated content on the web has massively increased through applications such as Facebook, Twitter, Flickr and LinkedIn, as well as commercial web sites. The value of such unbiased, real-time user-generated content has been shown to be tremendous, with applications in areas such as marketing, decision support systems, politics and public policy support. Due to the enormous amount of user content, it is a very difficult and challenging task to summarize information from online user content. Many natural language processing and information retrieval systems have been designed to automatically treat text and opinion utilizing subjectivity and sentiment analysis [1–4].

Different methods have been applied in sentiment classification tasks. These methods can be categorized into two main groups: lexicon-based and corpus-based [1,5,6]. Both the lexicon-based and corpus-based methods draw sentiment classification information from expert-annotated data sets. However, all of these sentiment classification resources are established in a limited number of languages, which leads to a resource imbalance between different languages. Most sentiment classification resources are written in English [7,8]. Furthermore, the manual construction of reliable sentiment resources is a very difficult and time-consuming task. Therefore, it would be advantageous to utilize labeled sentiment resources in one language (i.e., English) for sentiment classification in another language. This fantastic idea motivates an interesting research area called cross-lingual sentiment classification (CLSC). The most direct solution to this problem is to use machine translation systems to directly project the information from one language into another [7–12]. Most existing works in this area have applied machine translation systems to translate labelled training data from the source language into the target language and perform sentiment classification into the target language [13,14]. Other researchers have employed machine translation in another manner, translating unlabeled test data from the target language into the source language and performing the classification in the source language [7,11,15]. A limited number of research works have used both directions of translation to create two different views of the training and test data to compensate for some of the translation limitations [8,12,16,17].

The large gap between different languages occurs naturally. Every language has its unique linguistic terms and writing styles.

* Corresponding author.
E-mail addresses: zhpeng@sxu.edu.cn (P. Zhang), wsg@sxu.edu.cn (S. Wang), lidy@sxu.edu.cn (D. Li).

http://dx.doi.org/10.1016/j.knosys.2016.06.004
0950-7051/

Even when expressing similar idea, there can be a great disparity in the metaphor and vocabulary in contents of different languages, leading to a much smaller word and phrase intersection between translations and native expressions, as well as different semantic feature distributions between original language and target language contents. As a result, CLSC tasks cannot achieve performance comparable to that obtained for monolingual sentiment classification tasks. To alleviate the problem of this language gap, auxiliary unlabeled corpora or unlabeled parallel corpora are added into the training stage [8,18] to provide more bilingual word features. This strategy extends the training set and makes origin language and target language closer in the representation space. However, the complementary data brings in useful information as well as noise at the same time because it does not have exactly the same distribution as the training data and the test data. Especially when there is already a distribution disparity between the training set and the test set, noise of the complementary data may cause more adverse impacts than benefits. Additionally, the complementary data itself requires effort to be obtained, which imposes restrictions on the application of these methods.

Based on the above analysis, we try to overcome the difficulty of distribution disparity directly without any auxiliary samples. This paper proposes a novel CLSC strategy of similarity discovery plus training data adjustment (SD-TDA). In the similarity discovery phase, we set up an aligned-translation topic model to generate a bilingual concept representation space where the difference in content between samples from both origin and target languages can be measured through the topic distribution. As the aligned-translation topic model takes in co-occurrence information of terms in one language and between the original and target languages, the relevance of cross-lingual sentiment can be sharply enhanced. Then, in the training data adjustment phase, we set up a semi-supervised process to generate a training set suitable for the target test set. The generated training set is a part of the labeled original language samples that are similar to the reference samples. The reference samples are informative unlabeled target language samples that can be classified by the semi-supervised process with a high degree of confidence. The final classifier will be trained on the generated training set to fulfill the cross-lingual sentiment classification task. When these two steps work together, the sentiment similarity between languages can be maximized, while the distribution gap can be minimized. Generally, our strategy forms a new framework to fit the distribution disparity between the training set and the test set.

2. Related works

Cross-lingual sentiment classification. Cross-lingual sentiment classification is a type of text classification task. Bel et al. [19] first proposed the cross-lingual sentiment classification task, while earlier studies focused on cross-lingual information retrieval. Traditional cross-lingual classification and information retrieval tasks usually build up semantic mapping between languages based on resources such as bilingual lexicons or bilingual parallel corpora [20,21]. Based on these resources, a semantic mapping can be established by algorithms that joint synonymous features from different languages to archive a bilingual knowledge transformation between languages. The bilingual lexicons and bilingual parallel corpora contain human knowledge from experts who engineer the resources. Methods making use of these bilingual resources virtually use human knowledge to accomplish the cross-lingual classification tasks. Information from foreign language data can sometimes bring about a comprehensive understanding of certain topics. Some works use foreign language data to obtain more semantic information for a certain task. Wan [22] employed the combined utilization of Chinese and English lexicons to improve the sentiment classification performance of Chinese texts. Because the lexicons are essential resources, some works focus on automatically engineering bilingual lexicons. Aria et al. [23] designed a generative model to learn a bilingual lexicon from monolingual corpus. Huang et al. [24] constructed domain-specific sentiment lexicon based on constrained label propagation. Andres et al. [25] developed a novel method to choose the best dictionary for cross-language word sense disambiguation. As bilingual resources require costly human labor that increases the expense of labeled bilingual data sets, many methods employ automatic translation systems to translate data and resources as an alternative to human labor. One successful automatic translation system assistant strategy is the bilingual co-training framework [12], which sets up parallel classifiers on both the original language data view and the translation data view, and thus it can combine the original and target language sentiment classification results to conclude the final sentiment polarity of the unlabeled samples. Other approaches for cross-lingual tasks can also be applied in sentiment classification tasks. Cross-lingual structure correspondence learning (CSCL) presents another way to discover similarity between languages [11], where semantic similarity can be found up through a few pivot features. The above methods are based on the main idea that semantic corresponding relations across languages can transform sentiment classification information. Further studies have tried to invoke internal content structures across languages.

Topic models for multilingual tasks. Many topic models have been developed to discover latent topics underlying text contexts. Using topic modeling, documents are associated with a number of latent topics, which correspond to both document clusters and compact representations identified from a corpus. Each document is assigned to the topics with different weights, which specify both the degree of membership in the clusters as well as the coordinates of the document in the reduced dimension space. One basic topic model is the Latent Dirichlet Allocation (LDA) [26]. LDA includes a process for generating the topics in each document, thus greatly reducing the number of parameters to be learned for representation and providing a clearly defined probability for arbitrary documents. The LDA is unsupervised, where only the words in the documents are modelled in the generative process. In many applications, documents appear together with corresponding labels such as categorization labels or language labels. The supervised topic models add document labels into the topic modeling process to discover effective latent topics and predict labels for unlabeled documents [27]. Lexicons can cooperate with topic models to achieve aspect level sentiment classification [28]. Fu et al. [29] designed the dynamic non-parametric joint sentiment topic mixture model to detect and track dynamic sentiment and topics. Topic models for multilingual tasks are usually like the LDA and the supervised topic models. A multilingual topic model for unaligned text [30] is one way of modeling a multilingual corpus. It does not assume any explicit parallelism but instead discovers a parallelism at the vocabulary level. For document-aligned corpuses, the Bilingual Latent Dirichlet Allocation model (BiLDA) [31] provides a generative process taking into account bilingualism, and it is initially designed for parallel document pairs. A supervised topic model also can be applied in multilingual tasks to handle labeled corpuses [32]. In this paper, we follow the ideal of constructing latent topics across languages. The topic model we proposed in this paper focuses on modeling aligned translation data. The automatic translation system can provide parallel translation terms, but not as good as a parallel corpus annotated by human experts. We try to overcome this disadvantage by using a semi-supervised structure, which will be expounded upon in the following sections.

Semi-supervised learning for cross-lingual classification tasks. Semi-supervised learning is an effective machine learning framework that is designed to fit the occasion of lacking super-

vised information. In cross-lingual classification tasks, the training data and test data are from different languages, causing a semantic gap between the training data and test data. The semantic gap can be regarded as a lack of supervised information in the training data to train an effective classifier. Thus, the semi-supervised learning framework is applied in cross-lingual classification tasks to overcome the semantic gap. Usually, the semi-supervised learning framework grasps cross-lingual information from unlabeled auxiliary samples. One successful method is bilingual co-training [12], which selects high-confidence unlabeled samples by constructing parallel classifiers on both the original language and target language. Then, the high-confidence samples can grasp enough cross-lingual classification information. During the learning procedure, the quality of the selected unlabeled samples is essential to the final performance. Samples with more cross-lingual classification information should be selected by the semi-supervised learning framework. One feasible measurement to select high-quality samples is the similarity density degree [15,33], which draws upon a sample's k nearest neighbors mean similarity as a density measurement. The similarity density degree will select the samples most tightly surrounded by other samples and avoid selecting outlier samples. Thus, the selected samples are the most representative and informative samples. On the other hand, the difference between the instance and its nearest neighbor can also be used as a selection criterion for informative instances [34]. If the instances confidence is high but the confidence of its neighbor is low, it will contain more classification information. In the context of cross-lingual tasks, the quality measurement can be constructed on multi-language views [12,16,33,35]. In this paper, we define the sample quality criterion in concept representation space, where the samples context information quantity can be measured directly.

3. Basic framework

The cross-lingual sentiment classification aims at predicting sentiment labels of target language samples using labeled training samples in the original language. The cross-lingual sentiment classification is one type of text classification task based on corpuses, but with its own characteristic. Different from traditional text classification tasks, in cross-lingual tasks, words included in the training sample and test samples naturally have different characters. During the traditional text classification procedure, each sample is usually presented as a vector, namely, a bag-of-words model, where each component in the vector represents a single word. However, in cross-lingual tasks, samples from different languages share no characters in common, and this disables the bag-of-words model. Vector components from different languages are orthogonal, which means that no category information about the target language samples can be obtained from the training samples. To perform the cross-lingual sentiment classification, the samples in different languages should be first represented in the same semantic space. Thus, sentiment category information can be transmitted from the original language samples to the target language samples by a certain pattern recognition method through the cross-lingual semantic space. In practice, the sentiment category information is usually transmitted by a classifier, and this could cause another problem. The classifier assumes that the training samples and test samples follow the same distribution in the representation space. However, there is a language gap in cross-lingual tasks, as different languages have their own ways of expressing sentiments, which leads to a difference between the distributions of the training and test data sets, even in the semantic representation space. As a result, the classifier may not perform as well as expected. How to build a semantic representation space and how to overcome the language gap are the major difficulties in the cross-lingual sentiment classification task. Previous research has built cross-lingual semantic transmission methods based on the assumption that the original and target languages samples share part of their content. A semi-supervised framework can be applied to extend this cross-lingual semantic intersection until gaining enough sentiment classification information about the target language, just as the co-training framework did. During the semi-supervised procedure, auxiliary unlabeled samples are used to enhance the semantic relationship between languages. These auxiliary unlabeled samples need to be close in content to the target task data to introduce valuable information rather than noise. Some manual efforts therefore have to be made to obtain the auxiliary data. We push this semi-supervised framework idea further into a deeper thinking. We still follow the cross-lingual semantic intersection assumption. The semantic intersection of the original and target languages means that part of the training samples provides correct classification information, while other samples remain as noise. We can therefore treat the training samples separately as effective samples and noise samples. In the representation space, the distribution of the training data can be regarded as a mixture of noise and correct classification information. If we select effective samples out from the noise, we can obtain a correct cross-lingual sentiment classifier. This motivates the current paper to investigate the semantic intersection between languages and develop a method to identify effective training samples.

In this paper, we try to classify target language unlabeled samples by making use of original language training data and machine translation. To achieve this goal, we need to construct a semantic representation space that allows us to rank the effective training samples. A two-stage framework is proposed, shown in Fig. 1, for cross-lingual text sentiment classification tasks. The two-stage framework contains two sequential steps: the similarity discovery stage and the training data adjustment stage. The similarity discovery stage maximizes the semantic intersection between languages and the training data adjustment stage overcomes the language gap by filtering effective training samples to refine the training data set.

In the similarity discovery stage of the two-stage framework, the aligned-translation topic model is employed to extract the word co-occurrence relationship between the original and target languages from the aligned-translation labeled and unlabeled data. Those synonymy semantic concepts hidden in the aligned-translation data of the original and target languages are extracted as the latent topics. Based on these latent topics, the aggregation of some words from both original and target languages samples under the same topic can be found. The two groups of words separately from the original and target languages presented in a certain latent topic with higher co-occurrence probabilities can be regarded to play a similar role in expressing the topic. The role of the similarity discovery stage is to find out which words from the two different languages can express the same topic. Thus, the similarity expounded in the similarity discovery stage refers to the cross-language topic expressing similarity of two groups of words separately from two different languages.

After the representation space is optimized, the language gap becomes the major obstacle to the CLSC task. The language gap refers to the divergent ways of content expression in different languages. It comes from cultures, language styles and living habits, resulting in a phenomenon in which reviews from different languages focus on relative aspects and tend to describe things with various metaphors. The training data adjustment stage selects effective training samples by using a semi-supervised process. First, some informative unlabeled samples from the target language data set are found by several policies. Then, we predict the sentiment labels of these informative samples. The informative samples with high-confidence predicted labels will be used as a reference to update the training set. The new training set contains the part of the

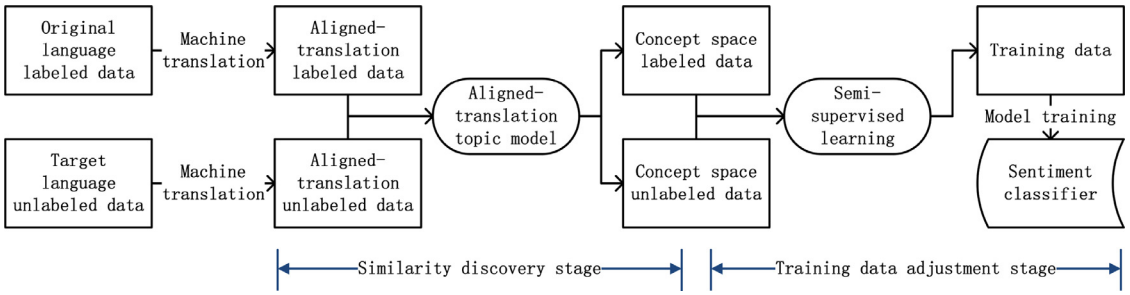


Fig. 1. Two-stage framework for modeling cross-lingual text sentiment classification. In the similarity discovery stage, both original and target languages data are firstly translated in the aligned-translation form. Then the aligned-translation topic model maps the aligned-translation data into the latent topic representation space. In training data adjustment stage, both labeled and unlabeled data re-expressed in the latent topic concept space are used in the semi-supervised learning procedure to iteratively refine the training data for constructing a target language sentiment classifier.

original labeled samples that are the closest to the reference samples. We repeat these steps until no more reference samples can be found, and the semi-supervised process produces an optimized training data set. The final classification model will be trained on the optimized training data set to complete the CLSC task.

The role of the training data adjustment stage playing in CLSC task is that it discoveries essential sentiment classification information really needed for the target language to refine the training samples. Owing to the language gap, training samples in the original language include some noise that contains error sentiment classification information about the target language. These noise samples need to be removed from the training set. To achieve this idea, the training data adjustment stage proposed in the current paper uses the informative and trustable samples from the target language as reference samples to rank the content of training samples. Labeled samples from the original language that express most similar content to the reference samples are regarded as high quality training samples. The training data adjustment stage presents a novel way to select the high quality training samples for the CLSC task.

4. Aligned-translation topic model (ATTM)

In this section, we introduce how to develop a topic model from aligned-translation data. The topic model could cluster relevant words to form a dense concept space to describe the content of the textual data. In cross-lingual sentiment tasks, we need a concept space holding topics from both original and target languages. The same concepts of synonymous content from different languages need to be clustered into relative topics. The cross-lingual synonym content relations usually have to be found from some congruent cross-lingual resources, such as parallel corpus or machine translation. In this paper, we refer to the aligned-translation as a cross-lingual semantic bridge. For the aligned-translation data, every sample has two views, the original language view and the target language view. The corresponding translation terms are placed in aligned positions. Although the aligned-translation texts are in different languages, they share the same meaning, so that we can assume that the original language view and the target language view are under the same topic distribution, while each language has its own word distribution. The synonymous topics can be grasped from the aligned-translation data. We define the ATTM based on this assumption.

4.1. Aligned-translation

Many online machine translation tools provide aligned-translation results. Commonly the webpage highlights corresponding terms of the original language and target language content. We use corresponding terms from Google Translation [1] as our aligned-translation data. Table 1 and Table 2 show examples of Chinese to English aligned-translation and English to Chinese aligned-translation, respectively.

As shown in the above tables, the machine translation has some syntax mistakes. In addition, the translation does not conform to the native language expression styles. However, these translation mistakes are sentence-level problems. The corresponding terms have been almost perfectly translated. Topic models cluster words in a text by their conditional distribution under topics. The sentence-level mistakes and syntax problems will not disturb the generation process of a certain word. In practice, the topic models treat the sentence as a bag of words, so that we can obtain effective latent topics on these cross-lingual aligned-translation terms while avoiding the adverse impact of translation mistakes. The aligned-translation provides two text views in two different languages with corresponding relations, which leads to a new need for double view topics modeling aligned-translation terms in a uniform way. We will introduce this type of topic model in the following section.

4.2. Definition of ATTM

The ATTM is similar to the LDA [26] model, which describes the generation process of corresponding terms in the aligned-translation data sets. Although the aligned-translation data sets are obtained from the machine translation service, the ATTM assumes that the bi-lingual view data can be generated by a hierarchical topic structure to learn a bi-lingual concept topic space. To introduce the ATTM, here we define some notations. The data set D contains M documents, denoted by $D = \{D_1, D_2, D_3, ..., D_M\}$, where D contains all original and target language samples. For each document, D_d contains N_d aligned-translation terms, which is defined as $D_d = \{< w_1^O, w_1^T >, < w_2^O, w_2^T >, < w_3^O, w_3^T > , ..., < w_{N_d}^O, w_{N_d}^T >\}$, where w_n^O represents the n-th term from the original language, and w_n^T represents n-th term from the target language.The generation process of the ATTM is shown in Fig. 2.

Based on the generation process of the ATTM, we design a probability structure shown in Fig. 3.

The dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed to be known and fixed. The word probabilities are parameterized by two matrixes, β^O and β^T, for the original and target languages, respectively; for now, we treat β^O and β^T as fixed quantities that are to be estimated.

A k-dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex (a k-vector θ lies in the $(k-1)$-simplex if

[1] http://translate.google.cn/

Table 1
Chinese to English aligned-translation

我住的是	新装修的	东楼,	房间挺	宽敞	感觉,	设施	也不错。
I live in a	newly renovated	East Building,	Room very	spacious	feeling,	Facilities	are good.

Table 2
English to Chinese aligned-translation

The breakfast	is	satisfactory and	hotel staff is	very polite.	The rooms are	very clean and	every day	of taken care.
早餐	是	令人满意的,	酒店的服务人员	很有礼貌。	房间	非常干净,	每天	照顾。

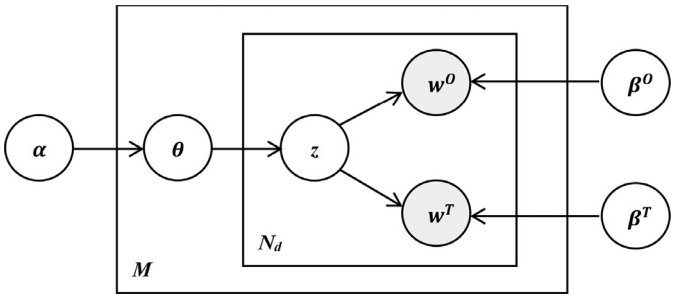


Fig. 2. Generation process of the ATTM.

(1) Choose $N_d \sim Poisson(\xi)$;
(2) Choose $\theta \sim Dir(\alpha)$;
(3) For each pair of aligned-translation terms $< w_n^O, w_n^T >$ in document set D:
  (a) Choose a topic $z \sim Multinomial(\theta)$;
  (b) Choose a conditional multinomial distribution $p(w_n^O, w_n^T|z, \beta^O, \beta^T)$ for $< w_n^O, w_n^T >$.

Fig. 3. Probability generation structure of the ATTM.

$\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$), and it has the following probability density on this simplex

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{1}$$

where parameter α is a k-vector with components $\alpha_i > 0$, and $\Gamma(x)$ is the gamma function. Given the parameters α, β^O and β^T, the joint distribution of a topic mixture θ, a set of N_d topics z, and a set of N_d aligned-translation terms is given by

$$p\left(\theta, z, w^O, w^T|\alpha, \beta^O, \beta^T\right) = p(\theta|\alpha) \prod_{n=1}^{N_d} p(z_n|\theta) p\left(w_n^O|z_n, \beta^O\right) p\left(w_n^T|z_n, \beta^T\right). \tag{2}$$

Integrating over θ and summing over z, we obtain the marginal distribution of a certain document D_d

$$p\left(D_d|\alpha, \beta^O, \beta^T\right) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z} p(z|\theta) p\left(w_n^O|z, \beta^O\right) p\left(w_n^T|z, \beta^T\right) \right) d\theta. \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of the corpus

$$p\left(D|\alpha, \beta^O, \beta^T\right) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z} p(z|\theta_d) p\left(w_n^O|z, \beta^O\right) p\left(w_n^T|z, \beta^T\right) \right) d\theta_d. \tag{4}$$

The conditional probability $p(\theta_d|\alpha)$ is the topic distribution in a certain document D_d.

4.3. Estimation of ATTM

We apply the variational method to estimate the parameters of the ATTM. By introducing variational parameters γ and φ, we assume the variational distribution as

$$q(\theta, z|\gamma, \varphi) = q(\theta|\gamma) \prod_{n=1}^{N_d} q(z_n|\varphi_n). \tag{5}$$

Thus, the optimized values of γ and φ can be estimated by an optimum condition

$$(\gamma^*, \varphi^*) = \arg\min D\left(q(\theta, z|\gamma, \varphi)||p(\theta, z|\alpha, \beta^O, \beta^T)\right). \tag{6}$$

We infer the updated equations of the variational parameters by applying the variational inference method upon the ATTM. The variational parameters can be updated as the following formulas

$$\varphi_{ni} \propto \beta^O_{iw_n^O} \cdot \beta^T_{iw_n^T} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j \right) \right) \tag{7}$$

$$\gamma = \alpha + \sum_{n=1}^{N_d} \varphi_n \tag{8}$$

The optimizing parameters (γ^*, φ^*) are document-specific. In particular, we view the Dirichlet parameters γ^* as providing a representation of a document in the topic simplex. For the fixed value of the variational parameters, φ^* minimizes the lower bound with respect to model parameters α, β^O and β^T. The conditional multinomial parameter β^O and β^T can be written as

$$\beta^O_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varphi_{dni} w^{Oj}_{dn} \tag{9}$$

$$\beta^T_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varphi_{dni} w^{Tj}_{dn}. \tag{10}$$

The Dirichlet parameter α can be implemented using an efficient Newton-Raphson method in the same way as the LDA's parameter estimation strategy. The model estimation procedure can be performed by the iterative algorithm shown in Algorithm 1.

The Algorithm 1 is called the alternating variational EM procedure, which repeats the E-step and the M-step until the ATTM achieves a stable status. The optimized variational parameter γ^* is viewed as the document concept representation in our further sample adjustment stage.

5. Semi-supervised training data adjustment

After the ATTM, both the original and target language samples are represented in the concept topic space. The ATTM clusters relative words from the original and target languages into corresponding topics, enhancing the coherence of the content concept. However the concept divergences of the original language and the

Algorithm 1 Model learning algorithm of the ATTM.

E-step: For each document D_d, compute the variational parameters$\{\gamma^*, \varphi^*\}$.

Initialize $\varphi_{ni}^0 := 1/k$, $n \in [1, N_d]$, $i \in [1, k]$;
Initialize $\gamma_i^0 := \alpha_i + N/k$, $i \in [1, k]$;
Repeat
  For $n = 1$ to N_d:
    for $i = 1$ to k:
      $\varphi_{ni}^{t+1} := \beta_{iw_n^O}^O \cdot \beta_{iw_n^T}^T \exp\left(\Psi\left(\gamma_i^t\right)\right)$;
    Normalize φ_n^{t+1} to sum to 1;
  $\gamma^{t+1} := \alpha + \sum_{n=1}^{N_d} \varphi_n^{t+1}$;
Until convergence.

M-step: Update model parameters α, β^O, β^T.

Update $\beta_{ij}^O = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varphi_{dni} w_{dn}^{Oj}$;
Normalize β_i^O to sum to 1;
Update $\beta_{ij}^T = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varphi_{dni} w_{dn}^{Tj}$;
Normalize β_i^T to sum to 1;
Update α with Newton-Raphson method.

target language cannot be eliminated by the similarity discovery strategy, for there are still topics on different aspects for different languages. The language gap still exists in the concept representation space, which means that different languages have divergences in understanding the positive and negative sentiments. Reviewers from different cultures focus on disparate topics and attributes of the same objects, and their representation also yields different metaphors. In the concept representation space, samples from different languages stretch in disparate directions, and the sentiment distributions of the training data and target language test data remain divergences. This leads to an essential need for training data adjustment.

Owing to the language gap, the training data cannot form an effective sentiment classifier. Only part of the labeled original language samples share the same content with the unlabeled target language samples in the concept representation space, and we call these labeled samples effective samples. The basic idea of the training data adjustment is to select the effective samples that can provide the most accurate sentiment information about the target language as a training set. Thus, we can learn a proper cross-lingual sentiment classifier on the well-chosen training set. To achieve this, some reference samples from the target languages are first selected by several criteria, and these reference samples have to contain the major content of the target language data set. The effective training samples are continuously filtered by a content similarity measurement based on the reference samples.

The whole procedure of training data adjustment is a type of semi-supervised learning framework. The framework puts effective samples into the training set at each iteration of the procedure. When no more reference samples can be found, the training set converges to the final training set. The semi-supervised training data adjustment procedure contains 3 major steps, shown in Fig. 4.

The quality measuring is first applied to the unlabeled samples to select informative unlabeled samples that contain the main content information about the unlabeled data set from the target language. Then, the part of the informative samples that can be classified with high confidence by the sentiment classifier is selected together with its predicted labels. These informative high-confidence samples are used as reference samples in the similarity measuring step, which is employed to select the training samples from the labeled original language samples. When the training samples are updated, the confidence measuring step and similarity measuring step iterate again. The newly selected informative high-confidence samples are added into the former set. When no more informative high-confidence samples are found, the semi-supervised alternating procedure ends. The final sentiment classifier will predict the final labels for the target language samples.

5.1. Sample quality measurement

The reference samples should be of high quality, which means that these samples contain main content and cover more topics in the concept space. Samples that cover more topics express more content of the data set. The complexity of the topic distribution in a sample can be used as a quality measurement to select the reference samples. We define a quality measurement on the topic distribution to indicate a value describing the complexity of a sample's topic distribution.

Definition 1. Let s be a sample with the representation vector [γ_1, γ_2, ..., γ_k] in a k-dimensional topic space. The quality $E(s)$ of sample s is defined as the Shannon entropy of the discrete probability distribution [$p(\gamma_1)$, $p(\gamma_2)$, ..., $p(\gamma_k)$],

$$E(s) = -\sum_{i=1}^{k} p(\gamma_i) \log_2 p(\gamma_i) \tag{11}$$

where $p(\gamma_i) = \frac{\gamma_i}{\sum_{j=1}^{k} \gamma_j}$.

The quality measurement $E(s)$ of a sample reaches its maximum value $log_2 k$ when $p(\gamma_i) = 1/k$ and the sample covers all topics with the same equal probability. When sample s contains only one topic, its quality measurement $E(s) = 0$. Samples with high quality measurement values provide more information about the target language, and we call these samples informative unlabeled samples. We set a minimum quality measurement threshold to select informative samples.

5.2. Sample confidence measurement

The adjustment of training samples needs to filter positive and negative samples separately. The reference samples also need to receive sentiment labels. And positive reference samples are used to filter positive training samples, and the negative reference samples the negative training samples. We employ a classifier to predict the discriminative probability of the selected informative samples. Each unlabeled sample s gets a probability belonging to positive category $p_+(s)$ and a probability belonging to negative category $p_-(s)$, restricted to $p_+(s) + p_-(s) = 1$, and we define the confidence measurement based on the predicted probabilities.

Definition 2. Let s be a sample and $p_+(s)$ and $p_-(s)$ be the probability values that s belongs to the positive and negative categories, respectively. The sample confidence measurement $C(s)$ of s is defined as

$$C(s) = |p_+(\mathrm{s}) - p_-(\mathrm{s})|. \tag{12}$$

The value of the sample confidence measurement $C(s)$ varies in the closed interval of [0, 1]. A higher value of the confidence measurement means that the sample is more clearly assigned to the positive or negative category. The predicted labels of the high confidence samples are reliable to be selected as references. We also set a minimum threshold to select high-confidence samples. For these high-confidence samples, we assume that their predicted labels are correct. These samples are then used as references in the following step.

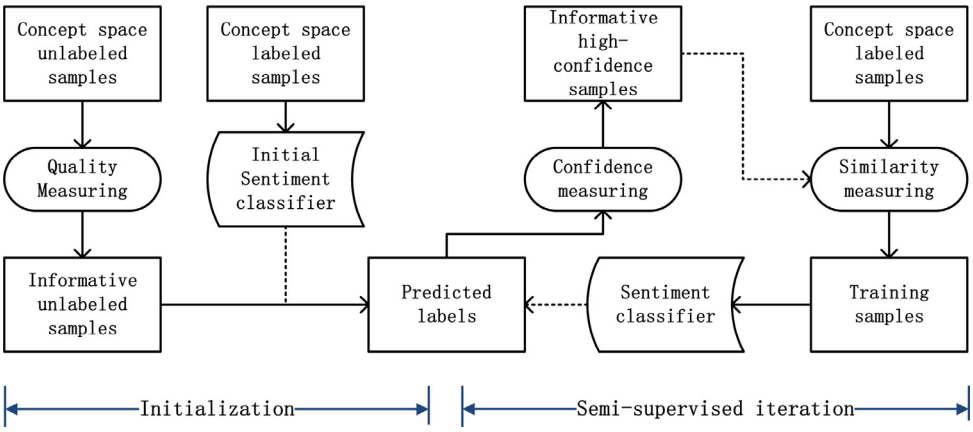


Fig. 4. Procedure of semi-supervised training data adjustment. Solid arrows mean data passing and dotted arrows mean being used in an operation. In initialization phase, all labeled samples are used to train the initial sentiment classifier. While in semi-supervised iteration phase, informative high-confidence samples are used to construct similarity measurement and then select training samples. The training samples will update the sentiment classifier and then update the predicted labels of informative unlabeled samples. The iterative procedure continues until no more samples can be selected as informative high-confidence samples.

5.3. Similarity measurement between a sample and a sample set

The similarity measurement is applied to the labeled original language samples to filtrate labeled training samples expressing sentiment information close to the target language data. We use the informative high-confidence samples from the target language as reference samples to select the labeled original language samples.

Definition 3. Let s and T be a sample and a sample set, respectively. The similarity measurement between s and T is defined as

$$S(s,T) = \exp\left(-\frac{1}{2}(s-\mu)^T \Sigma^{-1}(s-\mu)\right) \tag{13}$$

where $\mu = E[T]$ is the expectation of T, and Σ is the covariance matrix of T.

$$\Sigma = E_{t\in T}\left[(t-\mu)(t-\mu)^T\right] \tag{14}$$

The covariance matrix Σ contains the attribute correlativity of the sample set T. The similarity measurement $S(s, T)$ is normalized by Σ^{-1}, and the similarity measurement $S(s, T)$ takes the correlativity between attributes into consideration. The attributes of sample set T are latent topics based on the ATTM, which gives the similarity measurement $S(s, T)$ the ability to discover the semantic similarity of a certain sample s and a sample set T. The value of the similarity measurement $S(s, T)$ varies in the interval of (0, 1], monotonically decreasing. When a sample's content is close to the main content of T, the value of $S(s, T)$ tends to be close to 1. The basic form of this strategy comes from the Gaussian kernel function. Due to the reference samples, the target language semantic information can be added into the similarity measurement. Finally, we set a minimum proportion of original language labeled samples to be treated as a training set to learn the classifier.

5.4. Semi-supervised alternating algorithm

The whole procedure of the semi-supervised training samples adjustment method can be accomplished by an alternating algorithm shown as Algorithm 2.

After the above procedure (Algorithm 2), we obtain a cross-lingual sentiment classifier. Our approach is willing to find an exact sentiment classifier based on current labeled training samples rather than introducing redundant unlabeled samples into the learning stage. Although the unlabeled auxiliary data sets are proven useful to overcome the language gap, unnecessary information that is not relevant to the target language sentiment topics is imported to limit the cross-lingual sentiment classification performance. Our approach is proven reliable in the empirical evaluation.

Algorithm 2 Algorithm for adjusting semi-supervised training samples.

Require: Concept space unlabeled sample set U, concept space labeled sample set L, minimum quality threshold qlt, minimum confidence threshold cfd, and minimum proportion threshold of training samples ppt.

Ensure: Training set P and classifier C.

1: Initialize the classifier C with L;
2: Initialize the selected reference sets $T^+ = \emptyset$, $T^- = \emptyset$;
3: Initialize the training set $P = \emptyset$;
4: Calculate the quality for each sample in U, and obtain informative unlabeled samples $I = \{s | s \in U, E(s) > qlt\}$;
5: Repeat
6: For each sample s in I:
7: Predict sentiment label of s with C:
8: If $C(s) > cfd$ and s is predicted a positive label, then $T^+ = T^+ \cup \{s\}$;
9: If $C(s) > cfd$ and s is predicted a negative label, then $T^- = T^- \cup \{s\}$;
10: Calculate the similarity $S(s, T^+)$ for each positive sample s in L, and form a list SL^+ for these samples, ordered by decreasing similarity;
11: Calculate the similarity $S(s, T^-)$ for each negative sample s in L, and form a list SL^- for these samples, ordered by decreasing similarity;
12: Update $P = \{$Top ppt proportion of $SL^+\} \cup \{$Top ppt proportion of $SL^-\}$;
13: Update classifier C with P;
14: Until no more samples are added into T^+ and T^-.

6. Experiments and evaluations

In this section, we introduce the experiments to evaluate the proposed method. The experiments included test datasets in four different languages, and the training datasets were in Chinese. We tested the influence of the parameters in our method, and selected the best parameters to verify the effectiveness of the proposed method. We compared our experimental results with the

co-training method, transductive support vector machine, and the best performance of the COAE2014 tasks. The details of our experiments and evaluations are provided in the following sections.

6.1. Experimental setup

Here, we introduce the datasets and evaluation metrics employed in our experiments. A necessary data preprocessing step was applied to both the test sets and training sets.

Test Sets: To assess the performance of the proposed approach, we used the labeled data sets published by the COAE2014 (Chinese Opinion Analysis Evaluation 2014). The COAE2014 set up a public cross-lingual sentiment classification task that contained four target languages: German, English, French and Spanish. All of these data were hotel reviews collected from native websites. The German, French, and Spanish data sets included 2000 reviews (1000 positive reviews + 1000 negative reviews), respectively, and the English data set included 4000 reviews (2000 positive reviews + 2000 negative reviews).

Training Set: We employed a Chinese labeled data set as our training set. This training set was published by the Institute of Computing Technology of the Chinese Academy of Sciences, and it consisted of Chinese hotel reviews collected from native Chinese websites. The training set included 4000 reviews (2000 positive reviews + 2000 negative reviews).

Data preprocessing: Each Chinese review was translated into German, English, French and Spanish with the aligned-translation form. Each German, English, French and Spanish review was also translated into Chinese with the aligned-translation form. Therefore, every review received an aligned-translation form of the original and target languages.

Evaluation Metrics: We used the standard precision, recall and F-measure to measure the performance of the positive and negative classes, respectively, and used the accuracy metric to measure the overall performance of the system. The metrics were defined the same as in general text categorization.

6.2. Model parameter selection

The chief purpose of the proposed approach is to find a proper training set from the original language labeled samples. The ATTM provides a concept space where each sample receives a topic distribution structure representation. Later, the semi-supervised procedure filtrates samples with three measurements based on the concept space representation. These strategies provide us with insight into the mutual relevance relation between the original and target language samples. In our empirical study, we tested the proposed approach with a wide range of parameter values and discussed the influence of the parameters.

(1) Influence of the dimensionality k in ATTM

The ATTM grasps correlative topics from the aligned-translation data. The dimensionality k in the ATTM is the only parameter that needs manual setting. With the increment of k, the perplexity of the obtained sample concept representation decreases continuously. However, the performance of the cross-lingual sentiment classification task does not depend on the direct representation of the perplexity. To test the influence of the dimensionality k in the ATTM on the performance of the cross-lingual sentiment classification task, we set maximum values for the quality, similarity sample proportion and confidence thresholds, so that no reference samples can be selected and the performance only depends on the concept representation. All of the labeled samples of the original language were used as a training set. The classifier used in our experiments was the LibSVM[2] with a linear kernel. The performances of the ATTM in different languages are shown in Fig. 5(a).

As we can see from Fig. 5(a), the performance in different languages varied sharply with the increment of k. Upon comparing the accuracy of the ATTM in the four languages, English demonstrated superior performance to the other three languages. Because English is more similar to Chinese in sentence structures, grammar and phraseology than the other three languages, the performance between Chinese and English was better than that of the other languages.

Parameter k is an essential factor to the performance. Especially in this cross-lingual task, the parameter k has to fit the original language data set and the target language data set at the same time. Either of the original or target language data not being well represented in the concept topic space would lead to a greater language gap and a worse cross-lingual performance. Our evaluation data sets (both training and test data sets) were all native hotel reviews on different native topics and aspects, which could bring more complexity to the cross-lingual concept representation. Hence, as shown in Fig. 5(a), the performance on different values of k did not maintain a stable pattern. Fortunately, the proper value of k is no more than 50 for all four languages. This means that the cross-lingual sentiment semantic exists in a lower-dimension space and forms an empirical rule for the parameter setting.

(2) Influence of sample quality

To test the influence of the quality measurement, we chose the proper value of k for four languages (German 10, English 50, French 40, Spanish 40). Because the value of the quality measurement depends on the ATTM parameter k, we defined the value of the quality threshold in our experiments as follows.

$$qlt = m\log_2 k, m \in \{0.3, 0.4, 0.5, \ldots, 0.9\} \tag{15}$$

The confidence threshold and minimum proportion threshold were 0.6 and 0.3, respectively, so that the limited reference samples selected from the unlabeled data sets were sensitive to the quality threshold. We began the experiments for each language at $m = 0.3$, increasing m by 0.1 until no reference samples could be selected. The results are shown in Fig. 5(b).

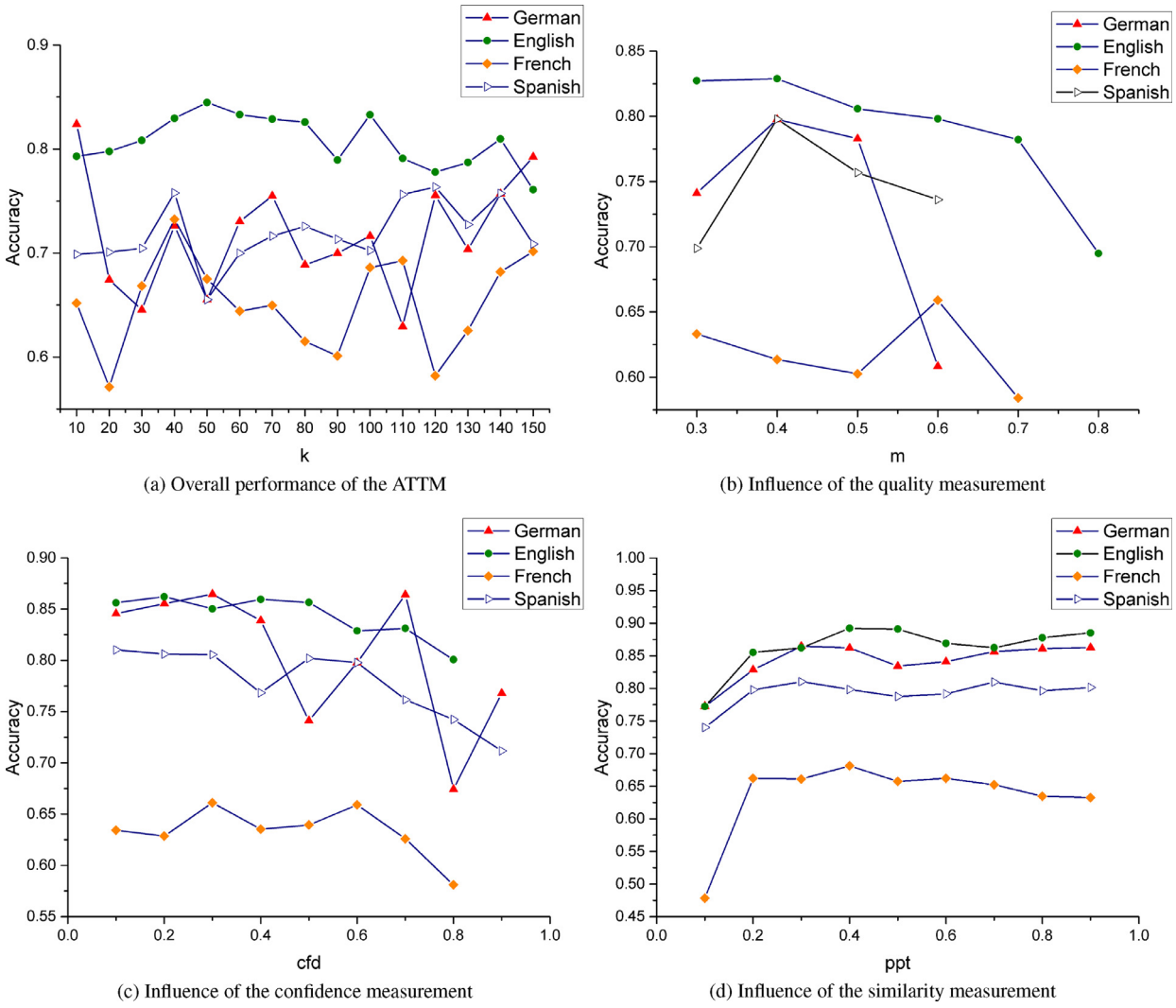
With the increment of m, the overall accuracy experienced a regular change. At the beginning, the accuracy increased as the selected reference samples obtained more accurate semantic information about the target data sets. When the overall accuracy reached its peak value, the quality threshold started to limit the number of reference samples. As a result, the performance decreased for a lack of reference samples. A basic principle that must be declared is that the quantity of reference samples in a certain data set is based on its properties. The quality measurement provides us with an efficient way to discover the reference samples by means of empirical tests.

(3) Influence of sample confidence

We tested the influence of the confidence measurement strategy based on the proper values of the quality measurement and the ATTM parameter. The proper values of the ATTM parameter k were the same as in the quality measurement experiment, and the proper values of the certainty measurement were 0.4, 0.4, 0.6 and 0.4 for German, English, French and Spanish, respectively. We set the confidence thresholds beginning at 0.1, and increased the thresholds by 0.1 until no reference samples could be selected. The results are shown in Fig. 5(c).

As shown in Fig. 5(c), the confidence measurement could improve the quality of the reference samples. With the increase of the confidence threshold, the overall accuracy first increased as the selected reference samples achieved more reliable predicted labels.

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/



(a) Overall performance of the ATTM

(b) Influence of the quality measurement

(c) Influence of the confidence measurement

(d) Influence of the similarity measurement

Fig. 5. Influence of parameters.

When the confidence threshold continued to increase, the number of reference samples decreased, which led to a performance reduction.

The previous quality measurement step already selected the informative samples as a candidate set. These informative samples obtained the major semantic content of the target data set. In addition, the informative samples were under a more balanced topic distribution. Thus, these samples would be more reliably predicted by the classifier, so that the proper confidence threshold would result in lower values.

(4) Influence of sample similarity

The similarity measurement is the last strategy used in the training data adjustment procedure. We added the proper values of the confidence threshold (German 0.3, English 0.2, French 0.3, Spanish 0.1) into the former proper parameter set for the ATTM, quality measurement and confidence measurement. We tested the proportion of the original language labeled training sample with a range of [0.1, 0.9] and incremented the proportion by 0.1. The results are shown in Fig. 5(d).

As we can see from Fig. 5(d), a part of the training set could achieve the best overall accuracy. In the beginning, the overall accuracy increased with the number of training samples. The growth of the training set introduced more sentiment information to the classifier. After the proportion threshold reached its peak accuracy, the training samples introduced more noise to the classifier, which led to a performance decrement. Although the overall accuracy might rise again as the proportion increases further, it could not reach the value at the proper threshold. This means that the noise of the training samples could not be eliminated by the scale of the training set.

6.3. Best model parameters

We have tested the influence of parameters in the proposed approach. The best parameters of the semi-supervised training data adjustment procedure are regrettably not simply a compilation of the proper parameters in each step. We have tested all combinations of the values of each parameter to find the best parameters for the proposed method. The results are shown in Table 3.

The proposed strategy can discover correlative sentiment information between different languages and eliminate noise of the training data as much as possible to form a proper cross-lingual sentiment classifier. As we can see from Table 3, fewer training samples and looser constraints are needed when the language gap is smaller, as in the Chinese to English sentiment classification task. For the other three languages, the language gap is more obvious, and more training samples and strict constraints are needed to select a proper training set. The proposed strategy is an effective

Table 3
Best parameters and best performance

Language	Best parameters				Positive			Negative			Total
	k	m	cfd	ppt	Recall	Precision	F-measure	Recall	Precision	F-measure	Accuracy
German	10	0.5	0.8	0.8	.921	.840	.879	.827	.914	.868	.874
English	50	0.4	0.2	0.4	.942	.856	.897	.843	.936	.887	.892
French	30	0.5	0.9	0.7	.853	.766	.807	.742	.836	.786	.797
Spanish	20	0.6	0.8	0.6	.908	.785	.842	.749	.890	.814	.829

Table 4
Performance of the co-training

Language	Positive			Negative			Total
	Recall	Precision	F-measure	Recall	Precision	F-measure	Accuracy
German	.848	.879	.863	.884	.855	.869	.866
English	.843	.845	.844	.847	.845	.846	.845
French	.791	.785	.788	.785	.791	.788	.788
Spanish	.837	.821	.829	.816	.832	.824	.826

Table 5
Mean performance of TSVM

Translation	Language	Positive			Negative			Total
		Recall	Precision	F-measure	Recall	Precision	F-measure	Accuracy
To original language	German	.870 ± .013	.838 ± .018	.853 ± .008	.833 ± .024	.867 ± .010	.849 ± .011	.851 ± .009
	English	.869 ± .020	.867 ± .009	.868 ± .009	.868 ± .012	.871 ± .016	.869 ± .007	.869 ± .008
	French	.753 ± .029	.755 ± .015	.754 ± .016	.758 ± .023	.757 ± .020	.757 ± .013	.756 ± .014
	Spanish	.808 ± .025	.814 ± .013	.811 ± .013	.814 ± .018	.809 ± .018	.811 ± .010	.811 ± .011
To target language	German	.919 ± .008	.751 ± .014	.826 ± .007	.699 ± .024	.898 ± .007	.785 ± .014	.808 ± .010
	English	.915 ± .015	.815 ± .007	.862 ± .005	.793 ± .013	.904 ± .014	.845 ± .005	.854 ± .005
	French	.892 ± .021	.682 ± .014	.773 ± .012	.587 ± .028	.847 ± .024	.693 ± .019	.739 ± .014
	Spanish	.877 ± .015	.756 ± .013	.812 ± .007	.715 ± .023	.852 ± .013	.777 ± .012	.796 ± .009

tool to analyze the relationship between different language data sets in sentiment classification tasks.

6.4. Method comparison

With the goal of testing the effectiveness of the proposed method, we employed a baseline for a comparison with our approach. The baseline methods include existing cross-lingual sentiment classification methods as well as public tasks. Then, we compared the proposed method and the baseline methods.

Baseline Methods: Many methods have been proposed for cross-lingual sentiment classification tasks. In this experiment, our approach was compared with the outstanding semi-supervised strategy of the co-training approach [12]. The baseline setting was the same as in Wan's co-training comparison setting ($I = 40$ and $p = n = 5$). Another general semi-supervised method, transductive support vector machine (TSVM) [3], was employed in the comparison. We performed the TSVM method in both original language and target language view. All words from training samples and auxiliary samples were used as features in the TSVM method. In addition, the best performance in the COAE2014 cross-lingual sentiment classification task was also compared with our approach. The classifier used in the co-training approach was also the LibSVM with a linear kernel. The performances of the baseline method and the best performance of COAE2014 [36] are shown in Tables 4, 5 and 6.

In the co-training method and TSVM method, unlabeled data are needed in the model learning stage. However, in our experiments, there were no auxiliary data sets. To complete the experiments, a part of the target language samples was used as auxiliary data. The co-training method could select unlabeled samples while the TSVM method needed manual unlabeled samples selection. We randomly selected unlabeled samples in TSVM method. The number of unlabeled samples in TSVM method was equal to that in the co-training method. We repeated the TSVM method 100 times, and the mean performance with its standard deviation is shown in Table 5. Above settings could introduce more information from the target language to improve the performance of co-training and TSVM method. Comparing Table 4 and 6, the co-training method achieves better results in German, and in the other target languages, the co-training method is close to the best performance of COAE2014. While in Table 5, the TSVM method performs better in the original view, and overall performance of the TSVM method is also close to the best performance of COAE2014. As the COAE2014 task did not put a limitation on the resources that could be used in the model training stage, we considered the co-training method and the TSVM method as acceptable baselines.

6.5. Discussion

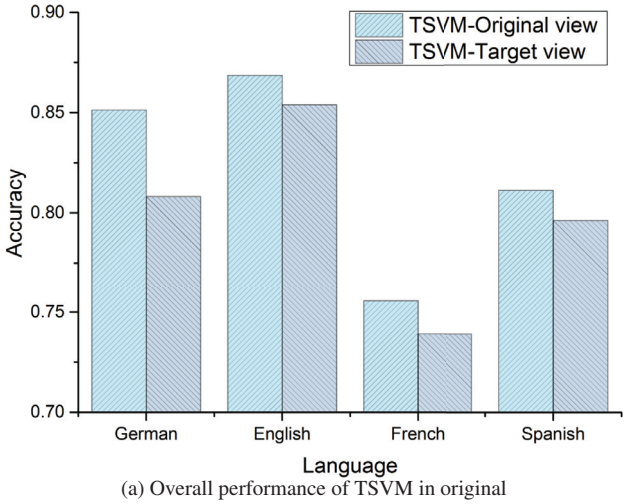
We put the positive F-measure, negative F-measure and overall accuracy of the proposed similarity discovery plus training data adjustment (SD-TDA) approach, co-training method, TSVM method (mean performance) and the best performance of COAE2014 for all four languages into a comparison, as shown in Fig. 6.

The language gap is obvious in our experimental results. Comparing the performance of the TSVM in original language view and target language view in Fig. 6(a), we can see the TSVM classifies target samples better in the original language view. In our experiments, the original language is Chinese. Chinese organizes sentences depending on semantic composition of words rather than syntactic structure. This characteristic gives Chinese words more semantic independence than other syntactic structure based

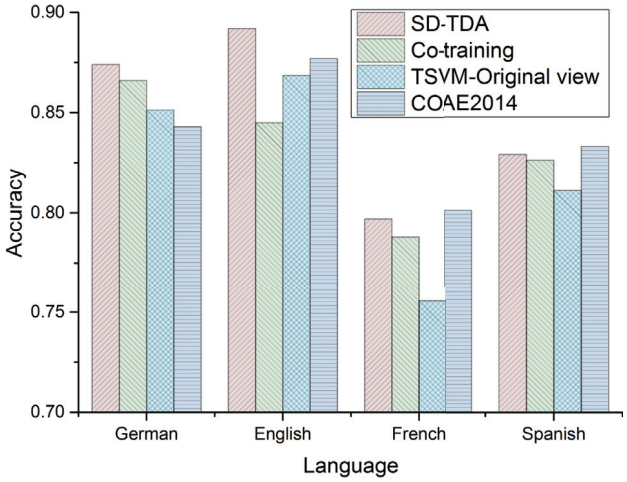
[3] http://www.cs.cornell.edu/People/tj/svm_light/

Table 6
Best performance of COAE2014

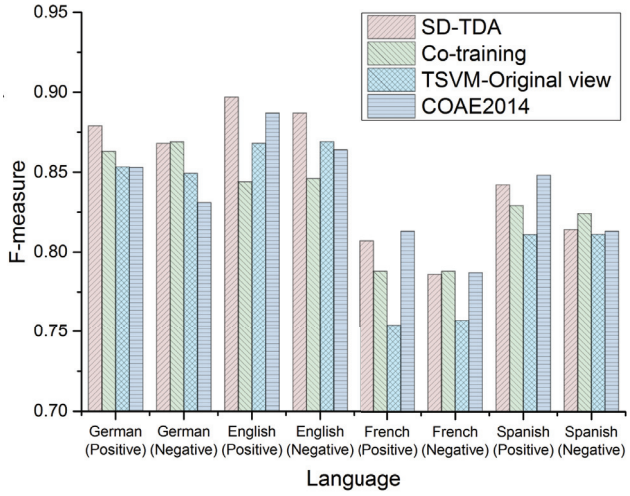
Language	Positive			Negative			Total
	Recall	Precision	F-measure	Recall	Precision	F-measure	Accuracy
German	.913	.801	.853	.773	.899	.831	.843
English	.968	.819	.887	.786	.960	.864	.877
French	.867	.766	.813	.735	.847	.787	.801
Spanish	.938	.775	.848	.727	.921	.813	.833



(a) Overall performance of TSVM in original language view and target language view



(b) Overall performance of the SD-TDA and the baselines



(c) Positive and negative F-measure of the SD-TDA and the baselines

Fig. 6. Method comparison.

languages. Thus the TSVM method did better in the original language view. Although the TSVM method takes in target language samples during the learning phase and optimizes the classification surface to fit the target language, the language gap cannot be merged on word level for words massively depending on each other in a syntactic structure based languages. When similarity between languages is stronger, like Chinese and English, the TSVM method would achieve better performance and the disparity between the original language view and the target language view would be smaller. To merge the language gap, we have to overcome the limitation of syntactic structure and build a concept representation space. This motivates the current paper to propose the ATTM.

The proposed training data adjustment approach is an effective semi-supervised method for CLSC tasks. As shown in Fig. 6(b), the SD-TDA method has attained or even exceeded the co-training method and the TSVM method. The semantic gap between languages makes the sentiment distributions over the training data and test data disagree with each other. As a result, training samples of inappropriate sentiment distribution to the test samples introduced noise information to the classifier. The co-training method can relieve the effect of the noise to a certain extent by appending auxiliary unlabeled samples. However, the noise of the training samples would never be removed by the co-training method, so the classification performance cannot be further improved. The same situation also occurs in the TSVM method. In the training sample adjustment phase, the reference samples filtered by quality measurement and confidence measurement grasp main semantics of target language. Based on these reference samples, the training set can be refined properly to exactly fit the

target language. Noise labeled samples would be left out of the training set so that the SD-TDA could achieve better performance.

The similarity between languages is the principal factor in the CLSC tasks. As presented in Fig. 6(b) and Fig. 6(c), the performance of Chinese to English is better than that of the other three languages. This is because Chinese is closer to English in structure and expression. The proposed approach exhibits more improvement in the Chinese to English task, which indicates that our method can divest semantic similarity between languages. Another phenomenon that we can see from Fig. 6(c) is that the performance of the positive aspect is better than that of the negative aspect. As a viewpoint of cognition, the concept of "*good*" is universally about a comfortable experience on a certain thing, such as "*Everything here is better than what we expected*" in the positive hotel reviews. However, the concept of "*bad*" is more likely to express a specific reason, such as "*The carpet is dirty with holes in it*" in negative hotel reviews. Thus, the similarity of the positive views is more coincident than that of the negative views in multi-lingual tasks, which leads to better performance in positive tasks.

The experiment shows that the proposed method has some performance differences in four target languages. We guess that the following reasons maybe cause the performance differences. The first cause should be language diversity, such as language style, phraseology, meaning of words and so on. For example, phenomena of polysemy and that a meaning can be expressed respectively by multiple words in different languages might cause some imperfections in the results of aligned-translation, and then affect the performance of sentiment classification for the target languages. Besides, the quality of aligned-translation is essential to our method. However, even though using the same aligned-translation tool, the quality of aligned-translation could present a big difference in different languages. For example, English is a common language widely used in the world. The popularity of English makes the machine translation systems perform better between English and other languages than between two non-English languages, which can also be confirmed in the result of our experiments. The quality of machine translation may be another important reason causing the performance difference of the proposed method. In addition, the content of training and test data will affect the performance of the proposed method as well. Data sets used in our experiments are native hotel reviews, and the contents of these reviews are not exactly about the same thing. Data sets that share closer contents may present better performance.

The SD-TDA expressed in the current paper leads to a new way of selecting a more exact training set for the cross-lingual sentiment classification task. Our approach can effectively avoid the influence of noisy samples and obtain better accuracy than the traditional semi-supervised method. The experiment shows that the proposed method is suitable for the cross-lingual sentiment classification task.

7. Conclusions and future work

This paper has proposed an effective method to fulfill the cross-lingual sentiment classification task. We concentrate on the primary obstacles of cross-lingual sentiment classification, which are the cross-lingual concept representation space construction and the language gap. The proposed similarity discovery plus training data adjustment framework includes two stages that separately form a concept representation space based on the cross-lingual word co-occurrence relations and overcome the language gap by finely filtering the training data to fit the target language. Thus, our framework has properly fit the situation of cross-lingual sentiment classification, and the obstacle of the language gap can be massively reduced. The experiments have shown that our framework achieves competitive performance with the co-training method, TSVM method and the best performance of COAE2014. Hence, the idea of similarity discovery plus training data adjustment is feasible for cross-lingual sentiment classification tasks. We can also systematically analyze the sentiment relationship and languages' semantic difference through the semi-supervised training data adjustment procedure. The basic idea that training data have to fit the target language samples can be extended to solve problems where discrepancies exist between the training samples and test samples. Our sample selection strategy can also be used to filter auxiliary samples in other semi-supervised frameworks.

The semantic similarity of different languages is a critical factor for cross-lingual tasks. In this paper, we discover cross-lingual semantic similarity based on the aligned-translation data. However, machine translation systems do not always perform as well as expected, and the generative process of topics ignores sentence structure; these shortcomings need to be improved upon. We have to continuously develop better cross-lingual concept representation methods by introducing more language analysis methods such as sentence syntactic structure or deep semantic concept structure to discover more latent sematic similarity relations between languages. When the representation space becomes better, semantic measurements become more accurate to distinguish differences of languages. Modeling cross-lingual deep latent semantic relations and constructing better multi-lingual representation spaces still need indispensable efforts in the future.

Acknowledgments

The authors would like to appreciate all anonymous reviewers for their valuable comments and suggestions which have significantly improved the quality and presentation of this paper. This work was supported by the National High-Tech Research and Development Program (863 Program) (2015AA011808); the National Natural Science Foundation of China (61432011, 61573231, 61175067, 61272095, U1435212, 41401521); the Shanxi Province Returned Overseas Research Project (2013-014); the Shanxi Province Science and Technology Basic Condition Platform Construction (2015091001-0102).

References

[1] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowl. Based Sys. 89 (2015) 14–46.
[2] A. Weichselbraun, S. Gindl, A. Scharl, Enriching semantic knowledge bases for opinion mining in big data applications, Knowl. Based Sys. 69 (2014) 78–85.
[3] A. Barrón-Cedeño, P. Gupta, P. Rosso, Methods for cross-language plagiarism detection, Knowl.Based Sys. 50 (2013) 211–217.
[4] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. trends inform.retriev. 2 (1-2) (2008) 1–135.
[5] R. Moraes, J.F. Valiati, W.P.G. Neto, Document-level sentiment classification: An empirical comparison between svm and ann, Expert Sys. Appl. 40 (2) (2013) 621–633.
[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. linguist. 37 (2) (2011) 267–307.
[7] M.-T. Martín-Valdivia, E. Martínez-Cámara, J.-M. Perea-Ortega, L.A. Ureña-López, Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches, Expert Sys. Appl. 40 (10) (2013) 3934–3942.
[8] X. Wan, Bilingual co-training for sentiment classification of chinese product reviews, Comput. linguist. 37 (3) (2011) 587–616.
[9] A. Balahur, R. Mihalcea, A. Montoyo, Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications, Comput. Speech Lang. 28 (1) (2014) 1–6.
[10] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual subjectivity analysis using machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 127–135.
[11] P. Prettenhofer, B. Stein, Cross-language text classification using structural correspondence learning, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 1118–1127.

[12] X. Wan, Co-training for cross-lingual sentiment classification, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Association for Computational Linguistics, 2009, pp. 235–243.
[13] A. Balahur, M. Turchi, Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis, Comput. Speech Lang. 28 (1) (2014) 56–75.
[14] C. Banea, R. Mihalcea, J. Wiebe, Multilingual subjectivity: are more languages better? in: Proceedings of the 23rd international conference on computational linguistics, Association for Computational Linguistics, 2010, pp. 28–36.
[15] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, Density based active self-training for cross-lingual sentiment classification, in: Advances in Computer Science and its Applications, Springer, 2014, pp. 1053–1059.
[16] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, Bi-view semi-supervised active learning for cross-lingual sentiment classification, Inform. Process. Manag. 50 (5) (2014) 718–732.
[17] J. Pan, G.-R. Xue, Y. Yu, Y. Wang, Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization, in: Advances in knowledge discovery and data mining, Springer, 2011, pp. 289–300.
[18] B. Lu, C. Tan, C. Cardie, B.K. Tsou, Joint bilingual sentiment classification with unlabeled parallel corpora, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, 2011, pp. 320–330.
[19] N. Bel, C.H. Koster, M. Villegas, Cross-lingual text categorization, in: Research and Advanced Technology for Digital Libraries, Springer, 2003, pp. 126–139.
[20] J.S. Olsson, D.W. Oard, J. Hajič, Cross-language text classification, in: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2005, pp. 645–646.
[21] V. Lavrenko, M. Choquette, W.B. Croft, Cross-lingual relevance models, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2002, pp. 175–182.
[22] X. Wan, Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 553–561.
[23] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, D. Klein, Learning bilingual lexicons from monolingual corpora., in: ACL, 2008, 2008, pp. 771–779.
[24] S. Huang, Z. Niu, C. Shi, Automatic construction of domain-specific sentiment lexicon based on constrained label propagation, Knowl. Based Sys. 56 (2014) 191–200.
[25] A. Duque, J. Martinez-Romo, L. Araujo, Choosing the best dictionary for cross-lingual word sense disambiguation, Knowl. Based Sys. 81 (2015) 65–75.
[26] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. mach. Learn. res. 3 (2003) 993–1022.
[27] J.D. Mcauliffe, D.M. Blei, Supervised topic models, in: Advances in neural information processing systems, 2008, pp. 121–128.
[28] F. Xianghua, L. Guo, G. Yanyan, W. Zhiqiang, Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon, Knowl.Based Sys. 37 (2013) 186–195.
[29] X. Fu, K. Yang, J.Z. Huang, L. Cui, Dynamic non-parametric joint sentiment topic mixture model, Knowl.Based Sys. 82 (2015) 102–114.
[30] J. Boyd-Graber, D.M. Blei, Multilingual topic models for unaligned text, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 75–82.
[31] I. Vulić, W. De Smet, M.-F. Moens, Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora, Inform. Retriev. 16 (3) (2013) 331–368.
[32] J. Boyd-Graber, P. Resnik, Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 45–55.
[33] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, H. Fujita, Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples, Inform. Sci. 317 (2015) 67–77.
[34] Y. Zhang, J. Wen, X. Wang, Z. Jiang, Semi-supervised learning combining co–training with active learning, Expert Syst. Appl. 41 (5) (2014) 2372–2378.
[35] Y. Liu, L. Dai, W. Zhou, H. Huang, Active learning for cross language text categorization, in: Advances in Knowledge Discovery and Data Mining, Springer, 2012, pp. 195–206.
[36] S. Tan, S. Wang, W. Xu, X. Yan, X. Liao, Overview of chinese opinion analysis evaluation 2014, in: Proceedings of COAE2014, Chinese Information Processing Society of China, 2014, pp. 5–25.