Cross-Lingual Adaptation using Structural Correspondence Learning

PETER PRETTENHOFER and BENNO STEIN
Bauhaus-Universität Weimar

Cross-lingual adaptation is a special case of domain adaptation and refers to the transfer of classification knowledge between two languages. In this article we describe an extension of Structural Correspondence Learning (SCL), a recently proposed algorithm for domain adaptation, for cross-lingual adaptation in the context of text classification. The proposed method uses unlabeled documents from both languages, along with a word translation oracle, to induce a cross-lingual representation that enables the transfer of classification knowledge from the source to the target language. The main advantages of this method over existing methods are resource efficiency and task specificity.

We conduct experiments in the area of cross-language topic and sentiment classification involving English as source language and German, French, and Japanese as target languages. The results show a significant improvement of the proposed method over a machine translation baseline, reducing the relative error due to cross-lingual adaptation by an average of 30% (topic classification) and 59% (sentiment classification). We further report on empirical analyses that reveal insights into the use of unlabeled data, the sensitivity with respect to important hyperparameters, and the nature of the induced cross-lingual word correspondences.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Cross-language text classification, cross-lingual adaptation, structural correspondence learning

ACM Reference Format:
Prettenhofer, P., Stein, B. 2010. Cross-lingual adaptation using structural correspondence learning. ACM Trans. Intell. Syst. Technol. 3, 1, Article 13 (October 2011), 22 pages.
DOI = 10.1145/2036264.2036277 http://doi.acm.org/10.1145/2036264.2036277

1. INTRODUCTION

Over the past two decades supervised machine learning methods have been successfully applied to many problems in natural language processing (e.g., named entity recognition, relation extraction, sentiment analysis) and information retrieval (e.g., text classification, information filtering). These methods, however, rely on large, annotated training corpora whose acquisition is time-consuming, costly, and inherently language-specific. As a consequence most of the available training corpora are in English only. Since an ever increasing fraction of the textual content available in digital

Author's addresses: Peter Prettenhofer, Bauhaus-Universität Weimar, 99421 Weimar, Germany, email: peter.prettenhofer@gmail.com; Benno Stein, Bauhaus-Universität Weimar, 99421 Weimar, Germany, email: benno.stein@uni-weimar.de.

DOI 10.1145/2036264.2036277 http://doi.acm.org/10.1145/2036264.2036277

form is written in languages other than English,[1] this fact limits the widespread application of state-of-the-art techniques from natural language processing (NLP) and information retrieval (IR). Technology for cross-lingual adaptation aims to overcome this problem by transferring the knowledge encoded within annotated (= labeled) data written in a source language to create a classifier for a different target language. Cross-lingual adaptation can thus be viewed as a special case of domain adaptation where each language acts as a separate domain.

In contrast to "classical" domain adaptation, cross-lingual adaptation is characterized by the fact that the two domains, i.e., the languages, have non-overlapping feature spaces, which has both theoretical and practical implications for domain adaptation. In classical domain adaptation—as well as in related problems such as covariate shift—overlapping feature spaces are implicitly presumed by the following or similar assumptions: (1) the existence of generalizable features, i.e., features which behave similarly in both domains [Jiang and Zhai 2007; Blitzer et al. 2006; Daume 2007], and, (2) the support of the test data distribution is contained in the support of the training data distribution [Bickel et al. 2009]. If, on the other hand, the feature sets are non-overlapping, one needs external knowledge to link features of the source domain and the target domain [Dai et al. 2008].

This article presents an approach for cross-lingual adaptation in the context of text classification: Cross-Language Structural Correspondence Learning (CL-SCL). CL-SCL uses unlabeled data from both languages along with external domain knowledge in the form of a word translation oracle to induce cross-lingual word correspondences. The approach is based on Structural Correspondence Learning (SCL), a recently proposed algorithm for domain adaptation in natural language processing [Blitzer et al. 2006; Blitzer et al. 2007].

Similar to SCL, CL-SCL induces correspondences among the words from both domains, i.e., languages, using a small number of so-called *pivots*. In CL-SCL, a pivot is a pair of words, $\{w_\mathcal{S}, w_\mathcal{T}\}$, from the source language $\mathcal{S}$ and the target language $\mathcal{T}$, which possess similar semantics. Testing the occurrence of $w_\mathcal{S}$ or $w_\mathcal{T}$ in a set of unlabeled documents from $\mathcal{S}$ and $\mathcal{T}$ yields two classes *across* these languages: one class contains the documents where either $w_\mathcal{S}$ or $w_\mathcal{T}$ occur, the other class contains the documents where neither $w_\mathcal{S}$ nor $w_\mathcal{T}$ occur. Ideally, a pivot splits the set of unlabeled documents with respect to the semantics that is associated with $\{w_\mathcal{S}, w_\mathcal{T}\}$. The correlation between $w_\mathcal{S}$ or $w_\mathcal{T}$ and other words w, $w \notin \{w_\mathcal{S}, w_\mathcal{T}\}$ is modeled by a linear classifier which then is used as a language-independent predictor for the two classes. A small number of pivots can capture a sufficiently large part of the correspondences between $\mathcal{S}$ and $\mathcal{T}$ in order to (1) construct a cross-lingual representation and (2) learn a classifier that operates on this representation. Several advantages follow from this approach:

— Task specificity. The approach induces task-specific word correspondences since it considers—during the pivot selection step—task-specific characteristics of language use.

— Efficiency in terms of linguistic resources. The approach uses unlabeled documents from both languages along with a small budget of calls (100 - 500) to a word translation oracle, instead of employing a parallel corpus or an extensive bilingual dictionary.

[1] This is especially the case for the World Wide Web where the number of Chinese speaking users has grown more than four times faster than the number of English speaking users in the last ten years (2000-2010). `http://www.internetworldstats.com/stats7.htm` (October 2010)

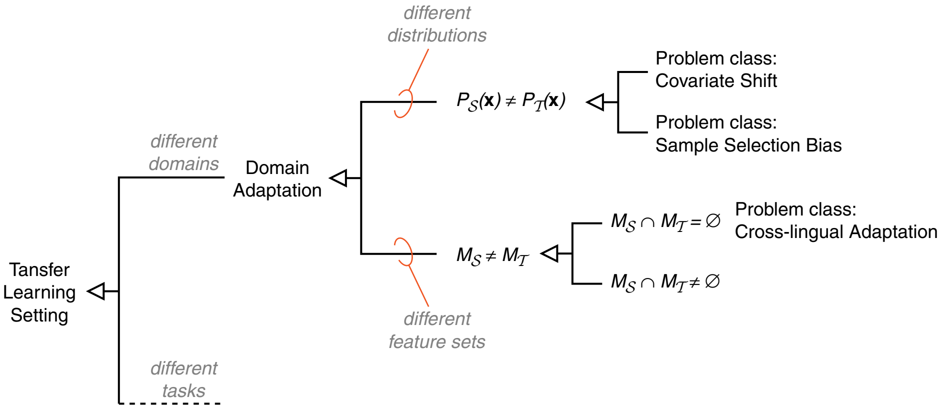

Fig. 1. A taxonomy of transfer learning settings, organized within the dimensions “domain” and “task”. The domain adaptation branch is unfolded.

— Efficiency in terms of computing resources. The approach solves the classification problem directly, instead of resorting to a more general and potentially much harder problem such as machine translation.

This article is an extended version of Prettenhofer and Stein [2010]. It includes a more detailed discussion of related work, computational considerations, and new experiments. In particular, we propose a novel strategy to obtain a sparse parameter matrix in the third step of CL-SCL, which leads to a significant improvement upon our previously reported results. This contribution has implications beyond CL-SCL, in particular for SCL [Blitzer et al. 2006] and related techniques such as Alternating Structural Optimization [Ando and Zhang 2005a].

The article is organized as follows: Section 2 discusses cross-lingual adaptation in the context of related work including domain adaptation and dataset shift. Section 3 introduces the problem of cross-language text classification, a special case of cross-lingual adaptation. Section 4 describes Cross-Language Structural Correspondence Learning, including computational considerations. Section 5 reports on the design and the results of experiments in the area of cross-language sentiment classification and topic classification. Section 6 concludes our work.

2. RELATED WORK

The idea to transfer knowledge from a source learning setting $\mathcal{S}$ to a different target learning setting $\mathcal{T}$ is an active field of research [Pan and Yang 2009], and Figure 1 organizes well-known problems of this field within a taxonomy. The taxonomy starts with the two most important determinants in a learning setting, namely, the *domain* and the *task*. A domain is defined by (1) a set of features M, (2) a space of possible feature vector realizations $\mathbf{x}$, which typically is the $\mathbb{R}^{|M|}$, and (3) a probability distribution $P(\mathbf{x})$ over the space of possible feature vector realizations.[2] A task specifies a set of labels corresponding to classes, typically $\{+1, -1\}$, along with a conditional distribution $P(y \mid \mathbf{x})$, with $y \in \{+1, -1\}$. Alternatively, a task can be specified by a sample $\{(\mathbf{x}, y) \mid \mathbf{x} \in \mathbb{R}^{|M|}, y \in \{+1, -1\}\}$. In Figure 1 the domain adaptation branch is unfolded since it is the focus of this article. The upper subbranch addresses problems where the feature sets are unchanged; without loss of generality $P_{\mathcal{S}}(\mathbf{x}) \neq P_{\mathcal{T}}(\mathbf{x})$ can also be presumed for problems in the lower subbranch “different feature sets”.

[2] If unambiguous the variables $\mathbf{x}$ and y may denote both a realization and a random variable.

2.1. Domain Adaptation

Domain adaptation refers to the problem of adapting a statistical classifier trained on data from one or more source domains to a different target domain. In the basic domain adaptation setting we are given labeled data from a source domain $\mathcal{S}$ and unlabeled data from the target domain $\mathcal{T}$, and the goal is to train a classifier for the target domain. Beyond this setting one can further distinguish whether a small amount of labeled data from the target domain is available [Daume 2007; Finkel and Manning 2009] or not [Blitzer et al. 2006; Jiang and Zhai 2007]. The latter setting is referred to as unsupervised domain adaptation.

Blitzer et al. [2006] propose an effective algorithm for unsupervised domain adaptation, called Structural Correspondence Learning. In a first step, SCL selects features that generalize across domains, which the authors call pivots. SCL then models the correlation between the pivots and all other features by training linear classifiers on the unlabeled data from both domains. This information is used to induce correspondences among features from the different domains and to learn a shared representation θ that is meaningful across both domains. This shared representation is appended to the original feature space, i.e., each instance is represented as $\mathbf{x}' = \left[\begin{smallmatrix}\mathbf{x}\\ \theta\mathbf{x}\end{smallmatrix}\right]$, a concatenation of the original and the shared representation. Concatenating the two feature spaces enables the classifier to exploit both generalizable features in the original representation as well as feature correspondences in the shared representation.

The major difference between CL-SCL and SCL refers to pivot selection and, in particular, the notion of a pivot. In SCL, pivots are features that generalize across domains. In cross-lingual adaptation the two domains have non-overlapping feature spaces, and hence one needs external knowledge to link the two feature spaces. In CL-SCL this external knowledge is represented by a feature translation oracle and pivots are pairs of features, one from each domain, that behave similarly in both domains. Furthermore, the original feature space cannot be used in cross-lingual adaptation, again due to non-overlapping feature spaces.[3]

SCL is related to the structural learning paradigm introduced by Ando and Zhang [2005a]. The basic idea of structural learning is to constrain the hypothesis space of a learning task by considering multiple different but related tasks on the same input space. Based on this paradigm Ando and Zhang [2005b] present a semi-supervised learning method, called Alternating Structural Optimization (ASO), which automatically generates related tasks from unlabeled data. These auxiliary tasks correspond to the pivots in SCL. In ASO, however, their purpose is to leverage information in unlabeled data in order to create a better feature space which improves the learning of the target task, especially in the presence of little labeled training data. The authors show that ASO delivers state-of-the-art performance for a variety of natural language processing tasks including named entity recognition and syntactic chunking. Quattoni et al. [2007] apply ASO to image classification in settings where little labeled data is given.

2.2. Dataset Shift

Traditional machine learning assumes that both training examples and test examples are drawn from identical distributions. This assumption is often violated in practice, for instance due to the irreproducibility of the test conditions during the training phase. Dataset shift refers to the general problem when the joint distribution of inputs (= feature vectors) and outputs (= labels) differs between training phase and test phase. The difference between dataset shift and domain adaptation is subtle; in

[3] The same idea can be used to combine the output of a machine translation system or a bilingual dictionary with the cross-lingual representation (cf. [Wei and Pal 2010; Margolis et al. 2010])

fact, both refer to the same underlying problem but emerge from the viewpoints of different research communities. Dataset shift is coined by the machine learning community and builds on prior work in statistics, especially the work on covariate shift [Shimodaira 2000] and sample selection bias [Cortes et al. 2008]. In contrast, domain adaptation originates from the natural language processing community. Most of the early work on domain adaptation focuses on the question of how to leverage "out-domain data" (= data associated with $\mathcal{S}$) effectively to learn a classifier when only little or no labeled "in-domain data" (= data associated with $\mathcal{T}$) is available. The latter emphasizes the relationship to semi-supervised learning—with the crucial difference that labeled data and unlabeled data stem from different distributions. Covariate shift can be considered as a special case of dataset shift that is closely related to unsupervised domain adaptation. Covariate shift is characterized by the fact that the underlying class conditional distribution between training phase and test phase is identical, i.e., $P_{\mathcal{S}}(y \,|\, \mathbf{x}) = P_{\mathcal{T}}(y \,|\, \mathbf{x})$, while the marginal distribution of the inputs (= covariates) differs, i.e., $P_{\mathcal{S}}(\mathbf{x}) \neq P_{\mathcal{T}}(\mathbf{x})$. An in-depth discussion of dataset shift is beyond the scope of this article, and the interested reader is referred to [Quionero-Candela et al. 2009].

2.3. Cross-Lingual Adaptation

Similar to domain adaptation, cross-lingual adaptation refers to the problem of adapting a statistical classifier trained on data from a source language $\mathcal{S}$ to a different target language $\mathcal{T}$. Examples include the adaptation of a named entity recognizer, a syntactic parser, or a relation extractor. The major characteristic of cross-lingual adaptation is the fact that the two "domains" have non-overlapping features sets, i.e., $M_{\mathcal{S}} \neq M_{\mathcal{T}}$. While cross-lingual adaptation has not received much attention in the natural language processing community,[4] a special case of cross-lingual adaptation recently gained considerable interest: cross-language text classification, which is also the focus of this article.

Bel et al. [2003] belong to the first who explicitly considered the problem of cross-language text classification. Their research, however, is predated by work in cross-language information retrieval (CLIR) where similar problems are addressed [Oard 1998]. Traditional approaches to cross-language text classification and CLIR use linguistic resources such as bilingual dictionaries or parallel corpora to induce correspondences between two languages [Lavrenko et al. 2002; Olsson et al. 2005]. Dumais et al. [1997] is considered as seminal work in CLIR: they propose a method which induces semantic correspondences between two languages by performing latent semantic analysis (LSA) on a parallel corpus. Li and Taylor [2007] improved upon this method by employing kernel canonical correlation analysis (CCA) instead of LSA. The major limitations of these approaches are their computational complexity and their dependence on parallel corpora, which are hard to obtain—especially for less resource-rich languages. Gliozzo and Strapparava [2005] circumvent the dependence on a parallel corpus by using so-called multilingual domain models, which can be acquired from comparable corpora in an unsupervised manner. In [Gliozzo and Strapparava 2006] they show for particular tasks that their approach can achieve a performance close to that of monolingual text classification.

Recent work in cross-language text classification focuses on the use of automatic machine translation technology. Most of these methods involve two steps: (1) translation of the documents into the source or the target language, and (2) dimensionality reduction or semi-supervised learning to reduce the noise introduced by the machine translation. Methods which follow this two-step approach include the EM-based approach

[4] There exists some relevant work on cross-lingual projection of annotations over aligned text [Riloff et al. 2002].

by Rigutini et al. [2005], the CCA approach by Fortuna and Shawe-Taylor [2005], the information bottleneck approach by Ling et al. [2008], and the co-training approach by Wan [2009]. The work of Wei and Pal [2010] and Margolis et al. [2010] is closely related to our work. Unlike our approach they also rely on automatic machine translation in a first step and then apply SCL to reduce the noise introduced by the machine translation.

3. CROSS-LANGUAGE TEXT CLASSIFICATION

In standard text classification, a document d is represented under the bag-of-words model as $|V|$-dimensional feature vector $\mathbf{x} \in \mathcal{X}$, where V, the vocabulary, denotes an ordered set of words, $x_i \in \mathbf{x}$ denotes the normalized frequency of word i in d, and $\mathcal{X}$ is an inner product space. $D_\mathcal{S}$ denotes the training set and comprises tuples of the form $(\mathbf{x}, y)$, which associate a feature vector $\mathbf{x} \in \mathcal{X}$ with a class label $y \in \mathcal{Y}$. For simplicity but without loss of generality we assume binary classification problems, $\mathcal{Y} = \{+1, -1\}$. The goal is to find a classifier $f : \mathcal{X} \to \mathcal{Y}$ that predicts the labels of new, previously unseen documents. In the following, we restrict ourselves to linear classifiers:

$$f(\mathbf{x}) = sign(\mathbf{w}^T\mathbf{x}), \tag{1}$$

where w is a weight vector that parameterizes the classifier and $[\cdot]^T$ denotes the matrix transpose. The computation of w from $D_\mathcal{S}$ is referred to as model estimation or training. A common choice for w is given by a vector w* that minimizes the regularized training error:

$$\mathbf{w}^* = \underset{\mathbf{w}\in\mathbb{R}^{|V|}}{\operatorname{argmin}} \sum_{(\mathbf{x},y)\in D_\mathcal{S}} L(y,\, \mathbf{w}^T\mathbf{x}) + \lambda R(\mathbf{w}) \tag{2}$$

L is a loss function that measures the effectiveness (= classification performance) of the classifier, R is a regularization term that penalizes model complexity, and λ is a non-negative hyperparameter that models the tradeoff between classification performance and model complexity. A common choice for R is L2-regularization, which imposes an L2-norm penalty on w, $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 = \frac{1}{2}\,\mathbf{w}^T\mathbf{w}$. Different choices for L entail different classifier types; e.g., when choosing the hinge loss function one obtains the popular Support Vector Machine classifier [Zhang 2004].

Standard text classification distinguishes between labeled (training) documents and unlabeled (test) documents. Cross-language text classification poses an extra constraint in that training documents and test documents are written in different languages. Here, the language of the training documents is referred to as source language $\mathcal{S}$, and the language of the test documents is referred to as target language $\mathcal{T}$. The vocabulary V divides into $V_\mathcal{S}$ and $V_\mathcal{T}$, called vocabulary of the source language and vocabulary of the target language, with $V_\mathcal{S} \cap V_\mathcal{T} = \emptyset$. I.e., documents from the training set and the test set map onto non-overlapping regions of the feature space. Thus, a linear classifier trained on $D_\mathcal{S}$ associates non-zero weights only with words from $V_\mathcal{S}$, which in turn means that it cannot be used to classify documents written in $\mathcal{T}$.

One way to overcome this "feature barrier" is to find a cross-lingual representation for documents written in $\mathcal{S}$ and $\mathcal{T}$ that transfers classification knowledge between the two languages. Intuitively, one can understand such a cross-lingual representation as a concept space that underlies both languages. In the following, we will use θ to denote a map that associates the original $|V|$-dimensional representation of a document d written in $\mathcal{S}$ or $\mathcal{T}$ with its k-dimensional cross-lingual representation, $\theta : \mathbb{R}^{|V|} \to \mathbb{R}^k$. Once such a mapping is found the cross-language text classification problem becomes a standard classification problem in the cross-lingual feature space. Note that the existing methods for cross-language text classification can be characterized by the way

θ is constructed. For instance, cross-language latent semantic indexing [Dumais et al. 1997] and cross-language explicit semantic analysis [Potthast et al. 2008] estimate θ using a parallel corpus. Other methods use linguistic resources such as a bilingual dictionary to obtain θ [Bel et al. 2003; Olsson et al. 2005; Wu et al. 2008].

4. CROSS-LANGUAGE STRUCTURAL CORRESPONDENCE LEARNING

We now present a method for learning a map θ by exploiting relations from unlabeled documents written in $\mathcal{S}$ and $\mathcal{T}$. The proposed method, which we call cross-language structural correspondence learning, CL-SCL, addresses the following learning setup (see also Figure 2):

(1) Given a set of labeled training documents $D_{\mathcal{S}}$ written in language $\mathcal{S}$, the goal is to create a text classifier for documents written in a different language $\mathcal{T}$. We refer to this classification task as the *target task*. An example for the target task is the determination of sentiment polarity, either positive or negative, of book reviews written in German ($\mathcal{T}$) given a set of training reviews written in English ($\mathcal{S}$).

(2) In addition to the labeled training documents $D_{\mathcal{S}}$ we have access to unlabeled documents $D_{\mathcal{S},u}$ and $D_{\mathcal{T},u}$ from both languages $\mathcal{S}$ and $\mathcal{T}$. Let D_u denote $D_{\mathcal{S},u} \cup D_{\mathcal{T},u}$.

(3) Finally, we are given a budget of calls to a word translation oracle (e.g., a domain expert) to map words in the source vocabulary $V_{\mathcal{S}}$ to their corresponding translations in the target vocabulary $V_{\mathcal{T}}$. For simplicity and without loss of applicability we can assume that the word translation oracle maps each word in $V_{\mathcal{S}}$ to exactly one word in $V_{\mathcal{T}}$.[5]

CL-SCL comprises three steps: In the first step, CL-SCL selects word pairs $\{w_{\mathcal{S}}, w_{\mathcal{T}}\}$, called pivots, where $w_{\mathcal{S}} \in V_{\mathcal{S}}$ and $w_{\mathcal{T}} \in V_{\mathcal{T}}$. Pivots have to satisfy the following conditions:

Confidence. Both words, $w_{\mathcal{S}}$ and $w_{\mathcal{T}}$, are predictive for the target task.
Support. Both words, $w_{\mathcal{S}}$ and $w_{\mathcal{T}}$, occur frequently in $D_{\mathcal{S},u}$ and $D_{\mathcal{T},u}$ respectively.

The confidence condition ensures that in the second step of CL-SCL only those correlations are modeled that are useful for discriminative learning. The support condition, on the other hand, ensures that these correlations can be estimated accurately. Considering our sentiment classification example, the word pair $\{\text{excellent}_{\mathcal{S}}, \text{exzellent}_{\mathcal{T}}\}$ satisfies both conditions: (1) the words are strong indicators of positive sentiment, and (2) the words occur frequently in book reviews from both languages. Note that the support of $w_{\mathcal{S}}$ and $w_{\mathcal{T}}$ can be determined from the unlabeled data D_u. The confidence, however, can only be determined for $w_{\mathcal{S}}$ since the setting gives us access to labeled data from $\mathcal{S}$ only.

We use the following heuristic to form an ordered set P of pivots: First, we choose a subset V_P from the source vocabulary $V_{\mathcal{S}}$, $|V_P| \ll |V_{\mathcal{S}}|$, which contains those words with the highest mutual information with respect to the class label of the target task in $D_{\mathcal{S}}$. Second, for each word $w_{\mathcal{S}} \in V_P$ we find its translation in the target vocabulary $V_{\mathcal{T}}$ by querying the translation oracle; we refer to the resulting set of word pairs as the candidate pivots P' :

$$P' = \{\{w_{\mathcal{S}}, \text{TRANSLATE}(w_{\mathcal{S}})\} \mid w_{\mathcal{S}} \in V_P\}$$

If the translation oracle fails to find a translation $w_{\mathcal{T}} \in V_{\mathcal{T}}$, $w_{\mathcal{S}}$ is discarded.

We enforce the support condition by eliminating in P' all candidate pivots $\{w_{\mathcal{S}}, w_{\mathcal{T}}\}$ where the document frequency of $w_{\mathcal{S}}$ in $D_{\mathcal{S},u}$ or of $w_{\mathcal{T}}$ in $D_{\mathcal{T},u}$ is smaller than some

[5] One can also consider a one-to-many mapping or a many-to-many mapping and adapt the definition of a pivot accordingly.

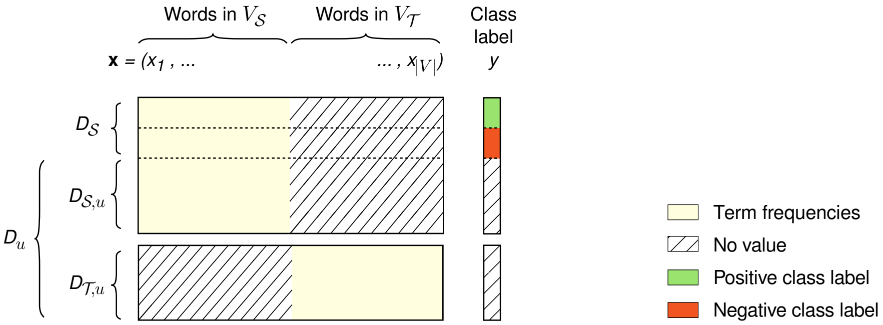

Fig. 2. The document sets underlying CL-SCL. The subscripts $_{\mathcal{S}}$, $_{\mathcal{T}}$, and $_u$ designate “source language”, “target language”, and “unlabeled”.

threshold ϕ:

$$P = \textsc{candidateElimination}(P', \phi)$$

Let m denote $|P|$, the number of pivots.

In the second step, CL-SCL models the correlations between each pivot $\{w_{\mathcal{S}}, w_{\mathcal{T}}\} \in P$ and all other words $w \in V \setminus \{w_{\mathcal{S}}, w_{\mathcal{T}}\}$. This is done by training classifiers that predict whether or not $w_{\mathcal{S}}$ or $w_{\mathcal{T}}$ occur in a document, based on the other words. For this purpose a training set D_l is created for each pivot $p_l \in P$:

$$D_l = \{(\textsc{mask}(\mathbf{x}, p_l),\ \textsc{in}(\mathbf{x}, p_l)) \mid \mathbf{x} \in D_u\}$$

$\textsc{mask}(\mathbf{x}, p_l)$ is a function that returns a copy of x where the components associated with the two words in p_l are set to zero, which is equivalent to removing these words from the feature space. $\textsc{in}(\mathbf{x}, p_l)$ returns $+1$ if one of the components of x associated with the words in p_l is non-zero and -1 otherwise. The support condition merely ensures that there is a sufficient number of positive training examples to accurately estimate the classifiers, i.e., to avoid cases of extreme class imbalance. For each D_l a linear classifier, characterized by the parameter vector w_l, is trained by minimizing Equation (2) on D_l. Recall that each training set D_l contains documents from both languages. Thus, for a pivot $p_l = \{w_{\mathcal{S}}, w_{\mathcal{T}}\}$ the vector $\mathbf{w}_l$ captures both the correlation between $w_{\mathcal{S}}$ and $V_{\mathcal{S}} \setminus \{w_{\mathcal{S}}\}$ and the correlation between $w_{\mathcal{T}}$ and $V_{\mathcal{T}} \setminus \{w_{\mathcal{T}}\}$.

In the third step, CL-SCL identifies correlations *across* pivots by computing the singular value decomposition of the $|V| \times m$-dimensional parameter matrix $\mathbf{W}$, $\mathbf{W} = [\mathbf{w}_1 \ \ldots \ \mathbf{w}_m]$:

$$\mathbf{U\Sigma V}^T \ = \ \mathrm{SVD}(\mathbf{W})$$

$\mathbf{W}$ encodes the correlation structure between pivot and non-pivot words in the form of multiple linear classifiers. I.e., the columns of $\mathbf{U}$ identify common substructures among these classifiers. Choosing the columns of $\mathbf{U}$ associated with the largest singular values yields the substructures that capture most of the correlation in $\mathbf{W}$. We define θ as those columns of $\mathbf{U}$ that are associated with the k largest singular values:

$$\theta = \mathbf{U}^T_{[1:k;\,1:|V|]}$$

Algorithm 1 summarizes the three steps of CL-SCL. At training and test time, we apply the projection θ to each input instance x. The vector $\mathbf{v}^*$ that minimizes the regularized training error for $D_{\mathcal{S}}$ in the projected space is defined as follows:

$$\mathbf{v}^* = \underset{\mathbf{v} \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{(\mathbf{x},y) \in D_{\mathcal{S}}} L(y,\, \mathbf{v}^T\theta\mathbf{x}) + \lambda R(\mathbf{v}) \qquad (3)$$

Algorithm 1 CL-SCL

```
Input:          $D_{\mathcal{S}}$   Labeled source data.
                $D_u$   Unlabeled data, $D_u = D_{\mathcal{S},u} \cup D_{\mathcal{T},u}$.
Parameters:     $m$     Number of pivots.
                $k$     Dimensionality of cross-lingual representation.
                $\lambda$     Regularization parameter.
                $\phi$     Support threshold.
Return value:   $\theta$     $k \times |V|$-dimensional matrix.

1. SELECTPIVOTS($D_{\mathcal{S}}, m$)
    $V_P = \text{MUTUALINFORMATION}(D_{\mathcal{S}})$
    $P' = \{\{w_{\mathcal{S}}, \text{TRANSLATE}(w_{\mathcal{S}})\} \mid w_{\mathcal{S}} \in V_P\}$
    $P = \text{CANDIDATEELIMINATION}(P', \phi)$

2. TRAINPIVOTCLASSIFIERS($D_u, P$)
    for $l = 1$ to $m$ do
        $D_l = \{(\text{MASK}(\mathbf{x}, p_l),\ \text{IN}(\mathbf{x}, p_l)) \mid \mathbf{x} \in D_u\}$
        $\mathbf{w}_l = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^{|V|}} \sum_{(\mathbf{x},y) \in D_l} L(y, \mathbf{w}^T\mathbf{x})) + \lambda R(\mathbf{w})$
    end for
    $\mathbf{W} = [\mathbf{w}_1 \ \dots \ \mathbf{w}_m]$

3. COMPUTESVD($\mathbf{W}, k$)
    $\mathbf{U\Sigma V}^T = \text{SVD}(\mathbf{W})$
    $\theta = \mathbf{U}^T_{[1:k;\,1:|V|]}$

return $\{\theta\}$
```

The resulting classifier, which will operate in the cross-lingual setting, is defined as follows:

$$f(\mathbf{x}) = sign(\mathbf{v}^{*T}\theta\mathbf{x})$$

4.1. Computational Considerations

Although the second step of CL-SCL involves the training of a fairly large number of linear classifiers, these classifiers can be learned efficiently due to (1) efficient learning algorithms for linear classifiers [Shalev-Shwartz et al. 2007] and (2) the fact that learning the pivot classifiers is an embarrassingly parallel problem. The computational bottleneck of the CL-SCL procedure is the SVD of the dense parameter matrix $\mathbf{W}$. Ando and Zhang [2005a] as well as Blitzer et al. [2007] propose to set negative entries in $\mathbf{W}$ to zero, in order to obtain a sparse matrix for which the SVD can be computed more efficiently [Berry 1992]. As a rationale for this step Ando and Zhang [2005a] claim that "the positive weights of a linear classifier are usually directly related to the target concept, while the negative components often yield much less specific information."

We propose a different strategy to obtain a sparse parameter matrix $\mathbf{W}$, namely to enforce sparse pivot classifiers $\mathbf{w}_l$ by employing a proper regularization term R in the second step of CL-SCL. A straight-forward solution is to use L1 regularization [Tibshirani 1996], which imposes an L1-norm penalty on w, $R(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{j=1}^{|V|} |w_j|$. This strategy recently gained much attention in the natural language processing community; Gao et al. [2007] show that L1 regularized models have similar predictive power to L2 regularized models while being much smaller at the same time—i.e., less parameters are non-zero.

L1 regularization, however, has properties which are inadequate in the context of SCL, in particular its handling of highly correlated features. Zou and Hastie [2005] show that if there is a subset of features among which the pairwise correlations are high, L1 regularization tends to select only one feature while pushing the other feature weights towards zero. This is certainly not desirable for SCL since it relies on the proper modeling of correlations in order to induce correspondences among features. L2 regularization, by contrast, exhibits a "grouping behavior", resulting in equal weights for correlated features. The SVD exploits this effect to find linear dependencies in the parameter matrix $\mathbf{W}$. The Elastic Net combines both properties, the sparsity property of L1 regularization and the grouping behavior of L2 regularization [Zou and Hastie 2005]. It is given by the convex combination of both norms:

$$R(\mathbf{w}) = \alpha\|\mathbf{w}\|_2^2 + (1-\alpha)\|\mathbf{w}\|_1, \tag{4}$$

where $\alpha \in [0;1]$ models the trade-off between grouping and sparsity. The Elastic Net is widely used in bioinformatics, in particular in the study of gene expression.

4.2. An Alternative View of CL-SCL

An alternative view of cross-language structural correspondence learning is provided by the framework of structural learning [Ando and Zhang 2005a]. The basic idea of structural learning is to constrain the hypothesis space, i.e., the space of possible weight vectors $\mathrm{w} \in \mathbb{R}^{|V|}$ of the target task, by considering multiple different but related auxiliary prediction tasks. Here, these auxiliary tasks are represented by the pivot classifiers, i.e., the columns of $\mathbf{W}$. Each column vector w_l can be considered as a linear classifier which performs well in both languages. θ^T defines the principal components of these bilingual classifiers; it characterizes what good bilingual classifiers are like. These principal components span a subspace of the parameter space which we regard as an approximation to the subspace of bilingual classifiers[6]. The basic idea is to find a classifier in this subspace which does well on $D_{\mathcal{S}}$, because—if θ^T is a good approximation to the subspace of bilingual classifiers—such a classifier will do equally well on the target data. Following Ando and Zhang [2005a] and Quattoni et al. [2007], we restrict the weight vector $\mathbf{w}^*$ for the target task to lie in the subspace defined by θ^T, $\mathbf{w}^* = \theta^T\mathbf{v}^*$, where $\mathbf{v}^*$ is defined as follows:

$$\mathbf{v}^* = \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^k} \sum_{(\mathbf{x},y)\in D_{\mathcal{S}}} L(y,\, (\theta^T\mathbf{v})^T\mathbf{x}) + \lambda R(\mathbf{v}) \tag{5}$$

Because of $(\theta^T\mathbf{v})^T = \mathbf{v}^T\theta$ this view of CL-SCL corresponds to the induction of a new feature space as given by Equation 3. Figure 3 illustrates the idea of the subspace constraint for $|V| = 3$, $m = 3$, and $k = 2$.

5. EXPERIMENTS

We evaluate CL-SCL for the tasks of cross-language sentiment classification and topic classification, using English as source language and German, French, and Japanese as target languages. We describe the experiment design, give implementation details, and present the evaluation results. Moreover, we give detailed analyses with respect to the nature of the induced cross-lingual correspondences, the use of unlabeled data, and important hyperparameters including the effect of different regularization terms for the pivot classifiers.

[6] Choosing the top-k principal components should be interpreted as a form of noise reduction since the pivot classifiers are corrupted with estimation errors.

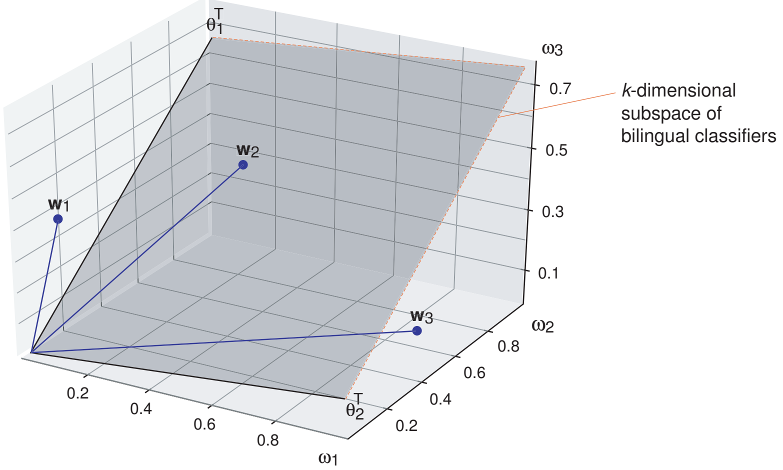

Fig. 3. Illustration of the subspace constraint for $|V| = 3$, $m = 3$, and $k = 2$. The plane spanned by θ_1^T and θ_2^T shows the subspace induced by the two principal components of $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]$. For the target task, we restrict the optimal weight vector $\mathbf{w}^*$ to lie in the 2-dimensional subspace defined by θ^T, i.e. we look for a bilingual classifier which does well on $D_{\mathcal{S}}$.

5.1. Datasets

We use the cross-lingual sentiment dataset provided by Prettenhofer and Stein [2010].[7] The dataset contains Amazon product reviews for the three product categories books, dvd, and music in the languages English, German, French and Japanese. Each document is labeled according to its sentiment polarity as either positive or negative. The documents in the dataset are organized by language and product category. There are three balanced disjoint sets of training, test, and unlabeled documents for each language-category pair; the respective set sizes are 2,000, 2,000, and 9,000-50,000. Similar to Prettenhofer and Stein [2010], each document d is represented as a normalized (unit length) feature vector $\mathbf{x}$ under a unigram bag-of-words model. Based on this dataset we create two tasks (see Table I for a summary statistics):

Sentiment Classification Task. For the task of cross-language sentiment classification we use the original partitioning of the cross-lingual sentiment dataset. Analogous to Prettenhofer and Stein [2010] we use English as source language, and German, French, and Japanese as target languages. For each of the nine target-language-category-combinations a sentiment classification task is created by taking the training set and the unlabeled set of the product category from $\mathcal{S}$ and the test set and the unlabeled set of the product category from $\mathcal{T}$.

Topic Classification Task. For the task of cross-language topic classification we discard the original sentiment labels and use the product category, i.e., books, dvd, and music as document label. Again we use English as source language and German, French, and Japanese as target languages. In contrast to the sentiment classification tasks, the classification of reviews according to product categories is a multi-class problem with mutually exclusive classes. Hence for each of the three target languages one cross-language topic classification task is created with the training set and the unlabeled set of all product categories from $\mathcal{S}$ and the test set and the unlabeled set of all product category from $\mathcal{T}$. For each of the three tasks we have 6,000 training and

[7] Available at `http://www.webis.de/research/corpora/webis-cls-10/`

Table I. Dataset statistics.

$\mathcal{T}$	Category	Unlabeled data		Labeled data		Vocabulary	
		$\lvert D_{\mathcal{S},u}\rvert$	$\lvert D_{\mathcal{T},u}\rvert$	$\lvert D_{\mathcal{S}}\rvert$	$\lvert D_{\mathcal{T}}\rvert$	$\lvert V_{\mathcal{S}}\rvert$	$\lvert V_{\mathcal{T}}\rvert$
German	books	50,000	50,000	2,000	2,000	64,682	108,573
	dvd	30,000	50,000	2,000	2,000	52,822	103,862
	music	25,000	50,000	2,000	2,000	41,306	99,287
French	books	50,000	32,000	2,000	2,000	64,682	55,016
	dvd	30,000	9,000	2,000	2,000	52,822	29,519
	music	25,000	16,000	2,000	2,000	41,306	42,097
Japanese	books	50,000	50,000	2,000	2,000	64,682	52,311
	dvd	30,000	50,000	2,000	2,000	52,822	54,533
	music	25,000	50,000	2,000	2,000	41,306	54,463
German	-	60,000	60,000	6,000	6,000	76,629	124,529
French	-	60,000	45,000	6,000	6,000	76,629	74,807
Japanese	-	60,000	60,000	6,000	6,000	76,629	64,050

Summary statistics for the nine cross-language sentiment classification tasks (first nine rows) and the three cross-language topic classification tasks (last three rows). All tasks use English as souce language $\mathcal{S}$. $|D_{\mathcal{S},u}|$ and $|D_{\mathcal{T},u}|$ give the number of unlabeled documents from $\mathcal{S}$ and $\mathcal{T}$; $|D_{\mathcal{S}}|$ and $|D_{\mathcal{T}}|$ give the number of training and test documents. All document sets are balanced.

6,000 test documents, containing a balanced number of examples. We set the number of unlabeled documents to 20,000 from each language-category pair.

5.2. Implementation

All experiments employ linear classifiers, which are trained by minimizing Equation (2) using stochastic gradient descent (SGD). We use the plain SGD algorithm as described by Zhang [2004] while adopting the learning rate schedule from PEGASOS [Shalev-Shwartz et al. 2007]. Similar to Blitzer et al. [2007] and Ando and Zhang [2005a], the modified Huber loss [Zhang 2004], a smoothed version of the hinge loss, is used as loss function L:

$$L(y, p) = \begin{cases} \max(0, 1 - py)^2, & \text{if } py \geq -1 \\ -4py, & \text{otherwise} \end{cases} \tag{6}$$

SGD and related methods based on stochastic approximation have been successfully applied to solve large-scale linear prediction problems in natural language processing and information retrieval [Zhang 2004; Shalev-Shwartz et al. 2007]. Their major advantages are efficiency and ease of implementation. SGD, however, cannot be applied directly in connection with L1 regularization (and thus the Elastic Net) due to the fluctuations of the approximated gradients. To overcome this problem different solutions have been proposed, such as methods based on truncated gradients [Langford et al. 2009; Tsuruoka et al. 2009] and projected gradients [Duchi et al. 2008]. In our experiments we resort to the truncated stochastic gradient algorithm proposed by Tsuruoka et al. [2009], which uses the cumulative L1 penalty to smooth out fluctuations in the approximated gradients. Note that Elastic Net regularization is applied for the pivot classifiers only; all other classifiers are trained using L2 regularization.

SGD receives two hyperparameters as input: the number of iterations T, and the regularization parameter λ. In our experiments T is always set to 10^6, which is about the number of iterations required for SGD to converge. For the target task, λ is determined by 3-fold cross-validation, testing for λ all values $10^{-i}, i \in [0;6]$. For the pivot prediction task, λ is set to the small value of 10^{-5}, in order to favor model accuracy over generalizability.

Since SGD is sensitive to feature scaling the projection $\theta\mathrm{x}$ is post-processed as follows: (1) Each feature of the cross-lingual representation is standardized to zero mean and unit variance, where mean and variance are estimated on $D_{\mathcal{S}} \cup D_u$. (2) The cross-lingual document representations are scaled by a constant β such that $|D_{\mathcal{S}}|^{-1} \sum_{\mathbf{x} \in D_{\mathcal{S}}} \|\beta\theta\mathbf{x}\| = 1$.

For multi-class classification the one-against-all strategy is applied. For multi-class problems, the subset V_P from the source vocabulary $V_{\mathcal{S}}$ is chosen as follows: (1) rank for each class the words according to mutual information with respect to all other classes, and (2) select the top ranked words from each ranking.

We follow Prettenhofer and Stein [2010] and use Google Translate as translation oracle,[8] which returns a single translation for each query word. In order to ensure the reproducibility of our results the cached translations provided by Prettenhofer and Stein [2010] are used whenever possible. Note that the word translation oracle operates context-free, which is clearly suboptimal, and we refrain from sanitizing the translations in order to demonstrate the robustness of CL-SCL with respect to translation noise.

For performance reasons, we implemented the SGD learning routine in C, the rest of code is written in Python.[9] The training of a single pivot classifier using Elastic Net regularization on 100.000 examples takes less than 3 seconds on a standard PC with an Intel Core2 Quad CPU with 2.83 GHz, exclusive the time to create the training examples. We train the pivot classifiers in parallel using a Hadoop cluster with 12 nodes of the above configuration. The training time for 450 pivot classifiers is less than two minutes, inclusive the time to create the training examples. For the SVD computation the Lanczos algorithm provided by SVDLIBC is employed.[10] The runtime of the SVD depends on the size of the vocabulary and the sparsity pattern of $\mathbf{W}$; in our experiments it is usually less than one minute.

5.3. Upper Bound and Baselines

To get an upper bound on the performance of a cross-language method we first consider the monolingual setting. For each task a linear classifier is learned on the training set of the target language and tested on the test set. The resulting accuracy scores are referred to as upper bound. This bound informs us about the expected performance on the target task if training data in the target language were available.

We choose two baselines to compare CL-SCL to other cross-language methods. The first baseline, CL-MT, is based on machine translation; it is a straightforward approach to cross-language text classification and has been used in a number of cross-language sentiment classification studies [Hiroshi et al. 2004; Bautin et al. 2008; Wan 2009]. CL-MT is determined as follows: (1) learn a linear classifier on the training data, (2) translate the test documents into the source language, and (3) predict the class label of the translated test documents. The translations of the test documents into the source language via Google Translate are provided by Prettenhofer and Stein [2010]. Note that CL-MT does not make use of unlabeled documents.

The second baseline, CL-Dict, uses the pivots obtained from the first step of CL-SCL as a bilingual dictionary to construct the mapping θ. This baseline is similar to the

[8] http://translate.google.com

[9] Our implementation is available at http://github.com/pprett/bolt

[10] http://tedlab.mit.edu/~dr/SVDLIBC/

Table II. Cross-language sentiment and topic classification results (absolute).

$\mathcal{T}$	Category	Upper Bound		CL-Dict		CL-MT		CL-SCL	
		μ	σ	μ	σ	μ	σ	μ	σ
German	books	83.79	±0.20	67.72	±0.08	79.68	±0.13	† **83.34**	±0.02
	dvd	81.78	±0.27	70.20	±0.22	77.92	±0.25	† **80.89**	±0.02
	music	82.80	±0.13	71.47	±0.20	77.22	±0.23	†† **82.90**	±0.00
French	books	83.92	±0.14	72.66	±0.10	80.76	±0.34	**81.27**	±0.08
	dvd	83.40	±0.28	72.55	±0.27	78.83	±0.19	**80.43**	±0.05
	music	86.09	±0.13	68.92	±0.50	75.78	±0.65	**78.05**	±0.06
Japanese	books	78.09	±0.14	58.79	±0.10	70.22	±0.27	†† **77.00**	±0.06
	dvd	81.56	±0.28	63.29	±0.25	71.30	±0.28	†† **76.37**	±0.05
	music	82.33	±0.13	65.27	±0.09	72.02	±0.29	†† **77.34**	±0.06
German	-	92.95	±0.11	87.99	±0.08	92.25	±0.07	**92.61**	±0.06
French	-	93.27	±0.07	88.00	±0.04	**90.58**	±0.17	90.57	±0.13
Japanese	-	89.43	±0.11	79.00	±0.09	82.14	±0.22	†† **85.03**	±0.10

Evaluation results for sentiment classification (first nine rows) and topic classification (last three rows) using English as source language $\mathcal{S}$. Accuracy scores (mean μ and standard deviation σ of 10 repetitions of SGD) on the test set of the target language $\mathcal{T}$ are reported. Statistical significance (McNemar) of CL-SCL is measured against CL-MT († indicates 0.05 and †† 0.001). For sentiment classification, CL-SCL uses $m = 450$, $k = 100$, $\phi = 30$, and $\alpha = 0.85$. For topic classification, CL-SCL uses $m = 250$, $k = 50$, $\phi = 50$, and $\alpha = 0.85$.

terminology translation approach in [Bel et al. 2003]. The difference between CL-Dict and CL-SCL informs us about the extent to which CL-SCL leverages information about word correlations from unlabeled data. In general we expect CL-Dict to be inferior to CL-MT since words are translated without taking their context into account. Note that CL-Dict uses unlabeled data only for pivot selection purposes.

5.4. Experimental Results

Table II contrasts the classification performance of CL-SCL with the upper bound and the baselines. Due to the inherent randomness of the training algorithm the accuracy scores are reported as mean μ and standard deviation σ of ten repetitions of SGD. McNemar's test is used to analyze whether or not the results of CL-SCL and CL-MT are statistically significant [Dietterich 1998]. Again, due to the randomness of the training algorithm, statistical significance is analyzed for each of the ten repetitions, whereas significance at a specific level is reported only if it applies to all repetitions.

Observe that the upper bound does not exhibit high variability across the three languages. For sentiment classification the average accuracy is about 82%, which is consistent with prior work on monolingual sentiment analysis [Pang et al. 2002; Blitzer et al. 2007]. For product category classification the average accuracy is in the low 90's, which is also consistent with prior work on monolingual product category classification [Crammer et al. 2009].

The performance of CL-MT differs considerably between the two European languages and Japanese: for Japanese, the averaged differences between the upper bound and CL-MT (9.5% for sentiment and 7.3% for topic) are much larger than for German and French (5.3% for sentiment and 1.7% for topic). This can be explained by the fact that machine translation works better for European than for Asian languages such as Japanese.

Table III. Cross-language sentiment and topic classification results (relative).

$\mathcal{T}$	Category	Upper Bound	CL-Dict	CL-MT	CL-SCL	
		μ	Δ_{UB}	Δ_{UB}	Δ_{UB}	RR[%]
German	books	83.79	16.07	4.11	**0.45**	89.05
	dvd	81.78	11.58	3.86	**0.89**	76.94
	music	82.80	11.33	5.58	**-0.10**	101.79
French	books	83.92	11.26	3.16	**2.65**	16.14
	dvd	83.40	10.85	4.57	**2.97**	35.01
	music	86.09	17.17	10.31	**8.04**	22.02
Japanese	books	78.09	19.30	7.87	**1.09**	86.15
	dvd	81.56	18.27	10.26	**5.19**	49.42
	music	82.33	17.06	10.31	**4.99**	51.60
German	-	92.95	4.96	0.70	**0.34**	51.43
French	-	93.27	5.27	**2.69**	2.70	-0.37
Japanese	-	89.43	10.43	7.29	**4.40**	39.64

Evaluation results of Table II relative to the upper bound. μ gives the accuracy score of the upper bound. Δ_{UB} gives the adaptation loss, i.e. the difference in accuracy to the upper bound. RR shows the relative reduction in error of CL-SCL over CL-MT.

CL-SCL receives four hyperparameters as input: the number of pivots m, the dimensionality of the cross-lingual representation k, the minimum support ϕ of a pivot in $D_{\mathcal{S},u}$ and $D_{\mathcal{T},u}$, and the Elastic Net coefficient α. For cross-language sentiment classification we use fixed values of $m = 450$, $k = 100$, $\phi = 30$, and $\alpha = 0.85$. For cross-language topic classification we found that smaller values of m and k work significantly better. The results for topic classification are obtained by using fixed values of $m = 250$, $k = 50$, $\phi = 50$, and $\alpha = 0.85$. The parameter settings have been optimized using the German book review task (sentiment) and the German topic task.

The results show that CL-SCL either outperforms or is at least competitive with CL-MT across all tasks. For German and Japanese sentiment classification we observe significant differences at a 0.05 and a 0.001 confidence level. For product category classification we observe significant differences only for Japanese (0.001 confidence level). Interestingly, for German music reviews, the accuracy of CL-SCL even surpasses the upper bound, which may be interpreted as a semi-supervised learning effect that stems from the massive use of unlabeled data.

Table III shows the classification performance relative to the upper bound, i.e., the loss due to cross-lingual adaptation. The rightmost column of Table III shows the relative reduction in error due to cross-lingual adaptation of CL-SCL over CL-MT; a relative reduction of 50% means that CL-SCL cuts the adaptation loss of CL-MT by 50%. CL-SCL reduces the relative error by an average of 59% (sentiment classification) and 30% (topic classification) over CL-MT. These results are a significant improvement upon our previously reported results in [Prettenhofer and Stein 2010], which is attributed to the use of a different regularization term for the pivot classifiers.

By contrasting the results of CL-SCL and CL-Dict one can also see that CL-SCL successfully exploits word correlations in the unlabeled data in order to create an effective cross-lingual representation. The poor performance of CL-Dict on certain tasks such as Japanese book reviews suggests that there is considerable noise in the translation oracle, and that CL-SCL is to a considerable extent robust to that noise.

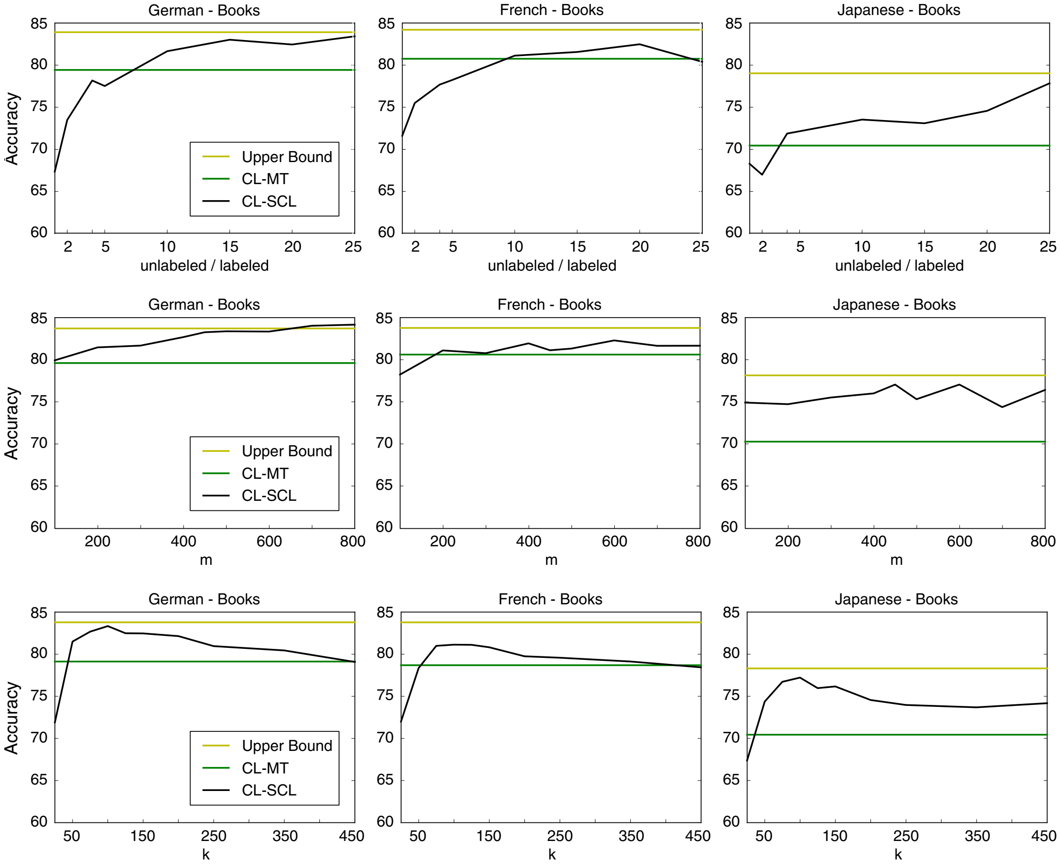

Fig. 4. Influence of unlabeled data and hyperparameters on the performance of CL-SCL. The rows show the performance of CL-SCL as a function of (1) the ratio between labeled and unlabeled documents, (2) the number of pivots m, and (3) the dimensionality of the cross-lingual representation k.

5.5. Sensitivity Analysis

This subsection analyzes the sensitivity of each of the four hyperparameters of CL-SCL in isolation while keeping the others fixed. If not specified otherwise, we use the same setting of the hyperparameters as in Table II.

Unlabeled Data. The first row of Figure 4 shows the performance of CL-SCL as a function of the ratio of labeled and unlabeled documents for sentiment classification of book reviews. A ratio of 1 means that $|D_{\mathcal{S},u}| = |D_{\mathcal{T},u}| = 2{,}000$, while a ratio of 25 corresponds to the setting of Table II. As expected, an increase in the number of unlabeled documents results in an improved performance. However, a saturation at a ratio of 10 can be observed across most tasks.

Number of Pivots. The second row shows the influence of the number of pivots m on the performance of CL-SCL. Compared to the size of the vocabularies $V_{\mathcal{S}}$ and $V_{\mathcal{T}}$, which is in 10^5 order of magnitude, the number of pivots is very small. The plots show that even a small number (100) of pivots captures a significant amount of the correspondence between $\mathcal{S}$ and $\mathcal{T}$.

Table IV. Effect of regularization.

$\mathcal{T}$	**Category**	**L2$^+$**		**L1**		**Elastic Net**		**L2**	
		μ	d[%]	μ	d[%]	μ	d[%]	μ	d[%]
German	books	79.50	17.88	82.45	1.24	83.34	11.02	**83.85**	66.28
	dvd	77.06	16.84	78.60	1.43	**80.89**	12.25	80.44	63.15
	music	77.60	16.00	81.41	1.72	82.90	13.92	**83.59**	61.54
French	books	79.02	16.50	80.75	1.87	81.27	14.13	**81.79**	65.26
	dvd	78.80	19.23	78.70	3.98	**80.43**	23.22	79.79	70.93
	music	77.72	16.70	77.32	3.72	78.05	21.60	**78.54**	65.92
Japanese	books	73.09	15.21	71.06	1.27	**77.00**	10.47	75.54	63.46
	dvd	71.10	14.86	75.75	1.48	76.37	11.84	**76.55**	62.12
	music	75.15	13.72	76.22	1.83	77.34	13.39	**77.79**	59.18
German	-	89.69	16.19	88.73	0.92	92.61	8.38	**92.75**	62.08
French	-	87.59	16.29	89.65	1.36	**90.57**	11.37	90.24	64.61
Japanese	-	82.83	16.71	84.26	1.23	**85.03**	10.15	84.43	66.01

The effect of different regularization terms on the performance of CL-SCL for cross-language sentiment (first nine rows) and topic classification (last three rows). Accuracy scores (mean μ of 10 repetitions of SGD) on the test set of the target language $\mathcal{T}$ are reported. d gives the density of the parameter matrix **W**, i.e., the number of non-zero entries divided by the total number of entries. **W** is $|V| \times 450$ where $|V|$ is in 10^5 orders of magnitude (see Table I for details). Elastic Net uses $\alpha = 0.85$.

Dimensionality of the Cross-Lingual Representation. The third row shows the influence of the dimensionality of the cross-lingual representation k on the performance of CL-SCL. Taking the top k left singular vectors can be regarded as a form of noise reduction, which has been introduced by (1) sub-optimal pivot selection and (2) estimation errors of w_l. Obviously the SVD is crucial to the success of CL-SCL if m is sufficiently large. Observe that the value of k is task-insensitive: a value of $50{<}k{<}150$ works equally well across all tasks.

Effect of Regularization. Table IV compares the effect of different strategies to obtain a sparse parameter matrix **W** on the performance of CL-SCL. The third column, L2$^+$, refers to the strategy in [Blitzer et al. 2006] and [Prettenhofer and Stein 2010] with L2 regularization and negative weights set to zero. Blitzer et al. [2006] claim that this strategy does not only reduce runtime and memory consumption but also improves the performance; however, we do not observe such a performance improvement over L2 in our experiments. The fourth column shows the performance of L1 regularization, which reduces the number of non-zero features compared to L2$^+$ from 16% to 2% on average. In Section 4.1 we argue that L1 regularization is not adequate due to its inadequate handling of highly correlated features and propose the Elastic Net penalty as an alternative. The empirical evidence supports this claim: Elastic Net regularization consistently outperforms both L2$^+$ and L1 regularization and is competitive with L2 while keeping the number of non-zero features low (15% on average). Elastic Net regularization adds an additional hyperparameter α that models the relative importance of L2 and L1 regularization. In the above experiments α is chosen such that the obtained density roughly equals the density of L2$^+$. A convenient property of the Elastic Net is that it encompasses L2 and L1 regularization as special cases. Thus, if m and $|V|$ are sufficiently small and a dense SVD is computationally feasible $\alpha = 1$ is optimal; otherwise, the optimal choice of α is governed by the computing resource.

Table V. Semantic and pragmatic correlations.

Pivot	English		German	
	Semantics	Pragmatics	Semantics	Pragmatics
$\{\text{beautiful}_{\mathcal{S}}, \text{schön}_{\mathcal{T}}\}$	amazing, beauty, lovely	picture, pattern, poetry, photographs, paintings	schöner (more beautiful), traurig (sad)	bilder (pictures), illustriert (illustrated)
$\{\text{boring}_{\mathcal{S}}, \text{langweilig}_{\mathcal{T}}\}$	plain, asleep, dry, long	characters, pages, story	langatmig (lengthy), einfach (plain), enttäuscht (disappointed)	charaktere (characters), handlung (plot), seiten (pages)

Semantic and pragmatic correlations identified for the two pivots $\{\text{beautiful}_{\mathcal{S}}, \text{schön}_{\mathcal{T}}\}$ and $\{\text{boring}_{\mathcal{S}}, \text{langweilig}_{\mathcal{T}}\}$ in English and German book reviews.

5.6. Interpretation of Results

The promising results of CL-SCL raise the question why CL-SCL is able to create more effective word correspondences than a state-of-the-art machine translation system. We argue that primarily responsible for the effectiveness of CL-SCL is its ability to induce task-specific word correspondences. Due to the use of task-specific, unlabeled data, relevant characteristics of task-specific language use are captured by the pivot classifiers. Table V exemplifies this claim with two pivots for German and English book reviews. The table show highly correlated words for the pivots $\{\text{beautiful}_{\mathcal{S}}, \text{schön}_{\mathcal{T}}\}$ and $\{\text{boring}_{\mathcal{S}}, \text{langweilig}_{\mathcal{T}}\}$, which are taken from the 50 highest positive entries in corresponding weight vectors. One can distinguish between (1) correlations that reflect similar meaning, such as “amazing”, “lovely”, or “plain”, and (2) correlations that reflect the pivot pragmatics with respect to the task, such as “picture”, “poetry”, or “pages”. Note in this respect that the authors of book reviews tend to use the word “beautiful” to refer to illustrations or to poetry, and that they use the word “pages” to indicate lengthy or boring books. We argue that while the first type of word correlations can be obtained by methods that operate on parallel corpora, the second correlation type requires an understanding of task-specific language use, which in general cannot be obtained from parallel corpora such as Europarl.

In order to gain further inside into the nature of the induced cross-lingual word correspondences Figure 5 shows two significant rows of θ, again for German and English book reviews.[11] Each row in θ projects all words in the vocabulary onto the real line. Words above the dashed line are from $V_{\mathcal{S}}$, words below the dashed line are from $V_{\mathcal{T}}$. Positive and negative values under these projections imply positive and negative sentiment respectively. The words in this illustration are selected as follows: (1) choose those two rows from θ which receive the highest (positive) weight from the final classifier, and (2) choose representative words from the 100 largest positive and negative entries in each row.

The first example clearly discriminates between well-written books which are well developed, organized, and researched, and books which are not, e.g. books with a questionable or far-fetched content. The second row, on the other hand, indicates whether or not a book had life-changing effects on the author of the book reviews. The word correspondences encoded within θ are indeed highly relevant to the task at hand, namely

[11] The visualization is inspired by Blitzer et al. [2006].

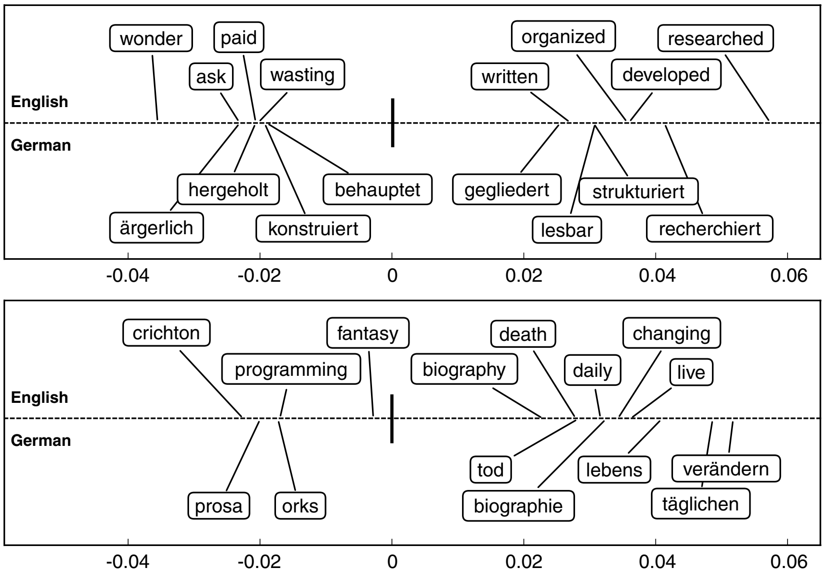

Fig. 5. Visualization of two significant rows of θ for English and German book reviews. Each row is a projection of the words onto the real line. Words on the left (right) indicate negative (positive) sentiment. Words close to each other behave similarly with regard to the classification task. The shown words are selected from the 100 largest positive and negative entries in each row.

discriminating between positive and negative book reviews. This is a major advantage of CL-SCL over existing concept-based approaches to cross-language text classification [Dumais et al. 1997; Gliozzo and Strapparava 2006; Li and Taylor 2007; Potthast et al. 2008] since they induce cross-lingual word correspondences without considering the target task.

6. CONCLUSIONS

This paper presents Cross-Language Structural Correspondence Learning, CL-SCL, as an effective technology for cross-lingual adaptation in the context of text classification. CL-SCL builds on Structural Correspondence Learning, a recently proposed algorithm for domain adaptation in natural language processing. CL-SCL uses unlabeled documents along with a word translation oracle to automatically induce task-specific, cross-lingual word correspondences.

We evaluated the approach for cross-language text classification, a special case of cross-lingual adaptation. The analysis covers performance and sensitivity issues in the context of sentiment and topic classification with English as source language and German, French, and Japanese as target languages. The results show a significant improvement of our approach over a machine translation baseline, reducing the relative error due to cross-lingual adaptation by an average of 59% (sentiment classification) and 30% (topic classification). Furthermore, the Elastic Net is proposed as an effective means to obtain a sparse parameter matrix, leading to a significant improvement upon our previously reported results [Prettenhofer and Stein 2010]. This technique has implications beyond CL-SCL, in particular for Structural Correspondence Learning [Blitzer et al. 2006] and Alternating Structural Optimization [Ando and Zhang 2005a].

Future work will concentrate on several open problems: while we introduced CL-SCL in the context of cross-language text classification, the method can be applied to any feature-based classifier if an appropriate feature translation oracle can be specified. As important target tasks we will investigate named entity recognition and relation extraction. These tasks require much richer feature representations than simple bag-of-words models. However, they also rely on word presence features which can be used to implement a simple word translation oracle analogous to the approach presented here. Furthermore, our approach makes no assumption about how the translation oracle is implemented: it could be a domain expert but also a heuristic based on lexical or phonetic similarity. Heuristics of this kind appear promising for technical domains such as patent databases or biomedical databases where similar technical expressions are used in both languages. This aspect opens a range of opportunities to further reduce the resource requirements of CL-SCL.

REFERENCES

ANDO, R. K. AND ZHANG, T. 2005a. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res. 6*, 1817–1853.

ANDO, R. K. AND ZHANG, T. 2005b. A high-performance semi-supervised learning method for text chunking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*. Association for Computational Linguistics, Morristown, NJ, USA, 1–9.

BAUTIN, M., VIJAYARENU, L., AND SKIENA, S. 2008. International sentiment analysis for news and blogs. In *Proceedings of ICWSM-08*. Seattle, 19–26.

BEL, N., KOSTER, C. H. A., AND VILLEGAS, M. 2003. Cross-lingual text categorization. In *Proceedings of ECDL-03*. Trondheim, 126–139.

BERRY, M. W. 1992. Large-scale sparse singular value computations. *International Journal of Supercomputer Applications 6,* 1, 13–49.

BICKEL, S., BRÜCKNER, M., AND SCHEFFER, T. 2009. Discriminative learning under covariate shift. *J. Mach. Learn. Res. 10*, 2137–2155.

BLITZER, J., DREDZE, M., AND PEREIRA, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL '07)*. Association for Computational Linguistics, Prague, Czech Republic, 440–447.

BLITZER, J., MCDONALD, R., AND PEREIRA, F. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*. Association for Computational Linguistics, Sydney, Australia, 120–128.

CORTES, C., MOHRI, M., RILEY, M., AND ROSTAMIZADEH, A. 2008. Sample selection bias correction theory. In *Algorithmic Learning Theory*, Y. Freund, L. Györfi, G. Turán, and T. Zeugmann, Eds. Lecture Notes in Computer Science, vol. 5254. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter 8, 38–53.

CRAMMER, K., DREDZE, M., AND KULESZA, A. 2009. Multi-class confidence weighted algorithms. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*. Association for Computational Linguistics, Morristown, NJ, USA, 496–504.

DAI, W., CHEN, Y., XUE, G.-R., YANG, Q., AND YU, Y. 2008. Translated learning: Transfer learning across different feature spaces. In *Advances in Neural Information Processing Systems 21*. MIT Press, 353–360.

DAUME, III, H. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL '07)*. Association for Computational Linguistics, Prague, Czech Republic, 256–263.

DIETTERICH, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation 10*, 1895–1923.

DUCHI, J., SHALEV-SHWARTZ, S., SINGER, Y., AND CHANDRA, T. 2008. Efficient projections onto the l_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, New York, NY, USA, 272–279.

DUMAIS, S. T., LETSCHE, T. A., LITTMAN, M. L., AND LANDAUER, T. K. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on CrossLanguage Text and Speech Retrieval. American Association for Artificial Intelligence.*

FINKEL, J. R. AND MANNING, C. D. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*. Association for Computational Linguistics, Morristown, NJ, USA, 602–610.

FORTUNA, B. AND SHAWE-TAYLOR, J. 2005. The use of machine translation tools for cross-lingual text mining. In *Workshop on Learning with Multiple Views, ICML '05*.

GAO, J., ANDREW, G., JOHNSON, M., AND TOUTANOVA, K. 2007. A comparative study of parameter estimation methods for statistical natural language processing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL '07)*. The Association for Computer Linguistics, Prague, Czech Republic, 824–831.

GLIOZZO, A. AND STRAPPARAVA, C. 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts (ParaText '05)*. Association for Computational Linguistics, Morristown, NJ, USA, 9–16.

GLIOZZO, A. AND STRAPPARAVA, C. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL '06)*. Association for Computational Linguistics, Morristown, NJ, USA, 553–560.

HIROSHI, K., TETSUYA, N., AND HIDEO, W. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th international conference on Computational Linguistics (ACL '04)*. Association for Computational Linguistics, Morristown, NJ, USA, 494+.

JIANG, J. AND ZHAI, C. 2007. A two-stage approach to domain adaptation for statistical classifiers. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM '07)*. ACM, New York, NY, USA, 401–410.

LANGFORD, J., LI, L., AND ZHANG, T. 2009. Sparse online learning via truncated gradient. *J. Mach. Learn. Res. 10*, 777–801.

LAVRENKO, V., CHOQUETTE, M., AND CROFT, W. B. 2002. Cross-lingual relevance models. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02)*. ACM, New York, NY, USA, 175–182.

LI, Y. AND TAYLOR, J. S. 2007. Advanced learning algorithms for cross-language patent retrieval and classification. *Inf. Process. Manage. 43,* 5, 1183–1199.

LING, X., XUE, G. R., DAI, W., JIANG, Y., YANG, Q., AND YU, Y. 2008. Can chinese web pages be classified with english data source? In *Proceeding of the 17th international conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 969–978.

MARGOLIS, A., LIVESCU, K., AND OSTENDORF, M. 2010. Domain adaptation with unlabeled data for dialog act tagging. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP '10)*. Association for Computational Linguistics, Uppsala, Sweden, 45–52.

OARD, D. W. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of AMTA-98*, D. Farwell, L. Gerber, E. H. Hovy, D. Farwell, L. Gerber, and E. H. Hovy, Eds. Lecture Notes in Computer Science, vol. 1529. Springer, 472–483.

OLSSON, J. S., OARD, D. W., AND HAJIČ, J. 2005. Cross-language text classification. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*. ACM, New York, NY, USA, 645–646.

PAN, S. J. AND YANG, Q. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering 99,* 1.

PANG, B., LEE, L., AND VAITHYANATHAN, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP '02)*. Association for Computational Linguistics, Morristown, NJ, USA, 79–86.

POTTHAST, M., STEIN, B., AND ANDERKA, M. 2008. A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*. Lecture Notes in Computer Science. Chapter 51, 522–530.

PRETTENHOFER, P. AND STEIN, B. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics (ACL '10)*. Association for Computational Linguistics, Uppsala, Sweden, 1118–1127.

QUATTONI, A., COLLINS, M., AND DARRELL, T. 2007. Learning visual representations using images with captions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 1–8.

QUIONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A., AND LAWRENCE, N. D. 2009. *Dataset Shift in Machine Learning*. The MIT Press.

RIGUTINI, L., MAGGINI, M., AND LIU, B. 2005. An EM based training algorithm for cross-language text categorization. *Web Intelligence, IEEE / WIC / ACM International Conference on 0*, 529–535.

RILOFF, E., SCHAFER, C., AND YAROWSKY, D. 2002. Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of the 19th international conference on Computational linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 1–7.

SHALEV-SHWARTZ, S., SINGER, Y., AND SREBRO, N. 2007. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning (ICML '07)*. ACM, New York, NY, USA, 807–814.

SHIMODAIRA, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference 90,* 2 (October), 227–244.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58,* 1, 267–288.

TSURUOKA, Y., TSUJII, J., AND ANANIADOU, S. 2009. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. 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*. Association for Computational Linguistics, Suntec, Singapore, 477–485.

WAN, X. 2009. 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*. Association for Computational Linguistics, Suntec, Singapore, 235–243.

WEI, B. AND PAL, C. 2010. Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Uppsala, Sweden, 258–262.

WU, K., WANG, X., AND LU, B.-L. 2008. Cross language text categorization using a bilingual lexicon. In *Proceedings of the Third International Joint Conference on Natural Language Processing*.

ZHANG, T. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning (ICML '04)*. ACM, New York, NY, USA, 116–124.

ZOU, H. AND HASTIE, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B 67,* 2 (April), 301–320.

Received June 2010; revised October 2010; accepted December 2010