Cross-Language Text Classification using Structural Correspondence Learning

Peter Prettenhofer and **Benno Stein**

Bauhaus-Universität Weimar
D-99421 Weimar, Germany
{peter.prettenhofer,benno.stein}@uni-weimar.de

Abstract

We present a new approach to cross-language text classification that builds on structural correspondence learning, a recently proposed theory for domain adaptation. The approach uses unlabeled documents, along with a simple word translation oracle, in order to induce task-specific, cross-lingual word correspondences. We report on analyses that reveal quantitative insights about the use of unlabeled data and the complexity of inter-language correspondence modeling.

We conduct experiments in the field of cross-language sentiment classification, employing English as source language, and German, French, and Japanese as target languages. The results are convincing; they demonstrate both the robustness and the competitiveness of the presented ideas.

1 Introduction

This paper deals with cross-language text classification problems. The solution of such problems requires the transfer of classification knowledge between two languages. Stated precisely: We are given a text classification task γ in a target language $\mathcal{T}$ for which no labeled documents are available. γ may be a spam filtering task, a topic categorization task, or a sentiment classification task. In addition, we are given *labeled* documents for the identical task in a different source language $\mathcal{S}$.

Such type of cross-language text classification problems are addressed by constructing a classifier $f_{\mathcal{S}}$ with training documents written in $\mathcal{S}$ and by applying $f_{\mathcal{S}}$ to unlabeled documents written in $\mathcal{T}$. For the application of $f_{\mathcal{S}}$ under language $\mathcal{T}$ different approaches are current practice: machine translation of unlabeled documents from $\mathcal{T}$ to $\mathcal{S}$, dictionary-based translation of unlabeled documents from $\mathcal{T}$ to $\mathcal{S}$, or language-independent concept modeling by means of comparable corpora. The mentioned approaches have their pros and cons, some of which are discussed below.

Here we propose a different approach to cross-language text classification which adopts ideas from the field of multi-task learning (Ando and Zhang, 2005a). Our approach builds upon structural correspondence learning, SCL, a recently proposed theory for domain adaptation in the field of natural language processing (Blitzer et al., 2006).

Similar to SCL, our approach induces correspondences among the words from both languages by means of a small number of so-called *pivots*. In our context 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 a 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 equivalence 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 equivalence classes. As we will see, 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 $f_{\mathcal{ST}}$ for the task γ that operates on this representation. Several advantages follow from our approach:

- Task specificity. The approach exploits the words' pragmatics since it considers—during the pivot selection step—task-specific characteristics of language use.

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1118–1127,
Uppsala, Sweden, 11-16 July 2010.

- Efficiency in terms of linguistic resources. The approach uses unlabeled documents from both languages along with a small number (100 - 500) of translated words, instead of employing a parallel corpus or an extensive bilingual dictionary.

- 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. Note that the use of such technology is prohibited in certain situations (market competitors) or restricted by environmental constraints (offline situations, high latency, bandwidth capacity).

Contributions Our contributions to the outlined field are threefold: First, the identification and utilization of the theory of SCL to cross-language text classification, which has, to the best of our knowledge, not been investigated before. Second, the further development and adaptation of SCL towards a technology that is competitive with the state-of-the-art in cross-language text classification. Third, an in-depth analysis with respect to important hyperparameters such as the ratio of labeled and unlabeled documents, the number of pivots, and the optimum dimensionality of the cross-lingual representation. In this connection we compile extensive corpora in the languages English, German, French, and Japanese, and for different sentiment classification tasks.

The paper is organized as follows: Section 2 surveys related work. Section 3 states the terminology for cross-language text classification. Section 4 describes our main contribution, a new approach to cross-language text classification based on structural correspondence learning. Section 5 presents experimental results in the context of cross-language sentiment classification.

2 Related Work

Cross-Language Text Classification 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) improve upon this method by employing kernel canonical correlation analysis, CCA, instead of LSA. The major limitation of these approaches is their computational complexity and, in particular, the dependence on a parallel corpus, which is 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 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).

Domain Adaptation Domain adaptation refers to the problem of adapting a statistical classifier trained on data from one (or more) source domains (e.g., newswire texts) to a different target domain (e.g., legal texts). In the basic domain adaptation setting we are given labeled data from the source domain and unlabeled data from the target domain, 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.

Note that, cross-language text classification can be cast as an unsupervised domain adaptation problem by considering each language as a separate domain. Blitzer et al. (2006) propose an effective algorithm for unsupervised domain adaptation, called structural correspondence learning. First, SCL identifies 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 that is meaningful across both domains. 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. Ando and Zhang (2005b) present a semi-supervised learning method based on this paradigm, which generates related tasks from unlabeled data. Quattoni et al. (2007) apply structural learning to image classification in settings where little labeled data is given.

3 Cross-Language Text Classification

This section introduces basic models and terminology.

In standard text classification, a document d is represented under the bag-of-words model as $|V|$-dimensional feature vector $\mathbf{x} \in X$, 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 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 $\mathbf{x} \in X$ with a class label $y \in Y$. The goal is to find a classifier $f : X \rightarrow Y$ that predicts the labels of new, previously unseen documents. Without loss of generality we restrict ourselves to binary classification problems and linear classifiers, i.e., $Y = \{+1, -1\}$ and $f(\mathbf{x}) = sign(\mathbf{w}^T\mathbf{x})$. $\mathbf{w}$ is a weight vector that parameterizes the classifier, $[\cdot]^T$ denotes the matrix transpose. The computation of $\mathbf{w}$ from $D_{\mathcal{S}}$ is referred to as model estimation or training. A common choice for $\mathbf{w}$ is given by a vector $\mathbf{w}^*$ that minimizes the regularized training error:

$$\mathbf{w}^* = \underset{\mathbf{w}\in\mathbf{R}^{|V|}}{\operatorname{argmin}} \sum_{(\mathbf{x},y)\in D_{\mathcal{S}}} L(y, \mathbf{w}^T\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \quad (1)$$

L is a loss function that measures the quality of the classifier, λ is a non-negative regularization parameter that penalizes model complexity, and $\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w}$. Different choices for L entail different classifier types; e.g., when choosing the hinge loss function for L 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 on two non-overlapping regions of the feature space. Thus, a linear classifier $f_{\mathcal{S}}$ trained on $D_{\mathcal{S}}$ associates non-zero weights only with words from $V_{\mathcal{S}}$, which in turn means that $f_{\mathcal{S}}$ 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 $\mathcal{T}$, which enables the transfer of 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 cross-lingual representation. Once such a mapping is found the cross-language text classification problem reduces to a standard classification problem in the cross-lingual 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).

4 Cross-Language Structural Correspondence Learning

We now present a novel 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 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 ($\mathcal{S}$).

- 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 $D_{\mathcal{S},u} \cup D_{\mathcal{T},u}$.

- 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 assume here that the word translation oracle maps each word in $V_{\mathcal{S}}$ to exactly one word in $V_{\mathcal{T}}$.

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 $\{\text{excellent}_{\mathcal{S}}, \text{excellent}_{\mathcal{T}}\}$ 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.

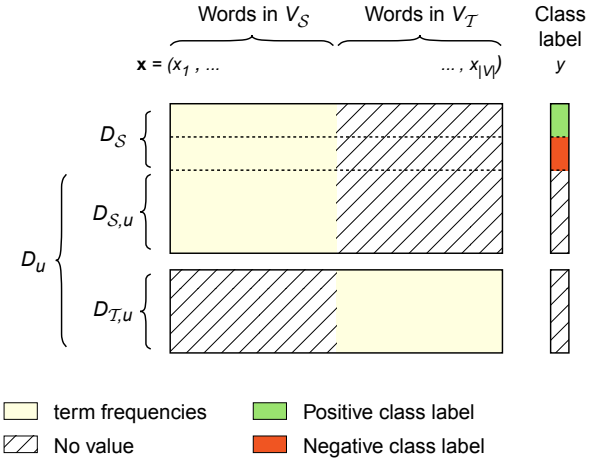

Figure 1: The document sets underlying CL-SCL. The subscripts $_{\mathcal{S}}$, $_{\mathcal{T}}$, and $_u$ designate "source language", "target language", and "unlabeled".

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':

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

We then 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 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 linear 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 $p_l \in P$:

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

MASK$(\mathbf{x}, p_l)$ is a function that returns a copy of $\mathbf{x}$ 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. IN$(\mathbf{x}, p_l)$ returns +1 if one of the components of $\mathbf{x}$ associated with the words in p_l is non-zero and -1 otherwise. For each D_l a linear classifier, characterized by the parameter vector $\mathbf{w}_l$, is trained by minimizing Equation (1) on D_l. Note 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 $\mathbf{W}$, $\mathbf{W} = \begin{bmatrix}\mathbf{w}_1 & \ldots & \mathbf{w}_m\end{bmatrix}$:

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

Recall that $\mathbf{W}$ encodes the correlation structure between pivot and non-pivot words in the form of multiple linear classifiers. Thus, the columns of $\mathbf{U}$ identify common substructures among these classifiers. Choosing the columns of $\mathbf{U}$ associated with the largest singular values yields those 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 $\mathbf{x}$. The vector $\mathbf{v}^*$ that minimizes the regularized training error for $D_\mathcal{S}$ in the projected space is defined as follows:

$$\mathbf{v}^* = \underset{\mathbf{v}\in\mathbf{R}^k}{\text{argmin}} \sum_{(\mathbf{x},y)\in D_\mathcal{S}} L(y,\, \mathbf{v}^T\theta\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{v}\|^2 \quad (2)$$

The resulting classifier $f_{\mathcal{ST}}$, which will operate in the cross-lingual setting, is defined as follows:

$$f_{\mathcal{ST}}(\mathbf{x}) = sign(\mathbf{v}^{*T}\theta\mathbf{x})$$

4.1 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, of the target task by considering multiple different but related prediction tasks. In our context these auxiliary tasks are represented by the pivot predictors, i.e., the columns of $\mathbf{W}$. Each column vector $\mathbf{w}_l$ can be considered as a linear classifier which performs well in both languages. I.e., we regard the column space of $\mathbf{W}$ as an approximation to the *subspace of bilingual classifiers*. By computing SVD$(\mathbf{W})$ one obtains a compact representation of this column space in the form of an orthonormal basis θ^T.

The subspace is used to constrain the learning of the target task by restricting the weight vector $\mathbf{w}$ to lie in the subspace defined by θ^T. Following Ando and Zhang (2005a) and Quattoni et al. (2007) we choose $\mathbf{w}$ for the target task to be $\mathbf{w}^* = \theta^T\mathbf{v}^*$, where $\mathbf{v}^*$ is defined as follows:

$$\mathbf{v}^* = \underset{\mathbf{v}\in\mathbf{R}^k}{\text{argmin}} \sum_{(\mathbf{x},y)\in D_\mathcal{S}} L(y,\, (\theta^T\mathbf{v})^T\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{v}\|^2 \quad (3)$$

Since $(\theta^T\mathbf{v})^T = \mathbf{v}^T\theta$ it follows that this view of CL-SCL corresponds to the induction of a new feature space given by Equation 2.

Algorithm 1 CL-SCL

Input: Labeled source data $D_\mathcal{S}$
Unlabeled data $D_u = D_{\mathcal{S},u} \cup D_{\mathcal{T},u}$

Parameters: m, k, λ, and ϕ

Output: $k \times |V|$-dimensional matrix θ

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. TRAINPIVOTPREDICTORS(D_u, P)
for $l = 1$ **to** m **do**
$D_l = \{(\text{MASK}(\mathbf{x}, p_l),\ \text{IN}(\mathbf{x}, p_l)) \mid \mathbf{x} \in D_u\}$
$\mathbf{w}_l = \underset{\mathbf{w}\in\mathbf{R}^{|V|}}{\text{argmin}} \sum_{(\mathbf{x},y)\in D_l} L(y, \mathbf{w}^T\mathbf{x})) + \frac{\lambda}{2}\|\mathbf{w}\|^2$
end for
$\mathbf{W} = \begin{bmatrix}\mathbf{w}_1 & \ldots & \mathbf{w}_m\end{bmatrix}$

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

output $\{\theta\}$

5 Experiments

We evaluate CL-SCL for the task of cross-language sentiment classification using English as source language and German, French, and Japanese as target languages. Special emphasis is put on corpus construction, determination of upper bounds and baselines, and a sensitivity analysis of important hyperparameters. All data described in the following is publicly available from our project website.[1]

5.1 Dataset and Preprocessing

We compiled a new dataset for cross-language sentiment classification by crawling product reviews from Amazon.{de | fr | co.jp}. The crawled part of the corpus contains more than 4 million reviews in the three languages German, French, and Japanese. The corpus is extended with English product reviews provided by Blitzer et al. (2007). Each review contains a category label, a title, the review text, and a rating of 1-5 stars. Following Blitzer et al. (2007) a review with >3 (<3) stars is labeled as positive (negative); other reviews are discarded. For each language the labeled reviews are grouped according to their category label, whereas we restrict our experiments to three categories: books, dvds, and music.

Since most of the crawled reviews are positive (80%), we decide to balance the number of positive and negative reviews. In this study, we are interested in whether the cross-lingual representation induced by CL-SCL captures the difference between positive and negative reviews; by balancing the reviews we ensure that the imbalance does not affect the learned model. Balancing is achieved by deleting reviews from the majority class uniformly at random for each language-specific category. The resulting sets are split into three disjoint, balanced sets, containing training documents, test documents, and unlabeled documents; the respective set sizes are 2,000, 2,000, and 9,000-50,000. See Table 1 for details.

For each of the nine target-language-category-combinations a text classification task is created by taking the training set of the product category in $\mathcal{S}$ and the test set of the same product category in $\mathcal{T}$. A document d is described as normalized feature vector $\mathbf{x}$ under a unigram bag-of-words document representation. The morphological analyzer MeCab is used for Japanese word segmentation.[2]

5.2 Implementation

Throughout the experiments linear classifiers are employed; they are trained by minimizing Equation (1), using a stochastic gradient descent (SGD) algorithm. In particular, the learning rate schedule from PEGASOS is adopted (Shalev-Shwartz et al., 2007), and the modified Huber loss, introduced by Zhang (2004), is chosen as loss function L.[3]

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.

The computational bottleneck of CL-SCL is the SVD of the dense parameter matrix $\mathbf{W}$. Here we follow Blitzer et al. (2006) and set the negative values in $\mathbf{W}$ to zero, which yields a sparse representation. For the SVD computation the Lanczos algorithm provided by SVDLIBC is employed.[4] We investigated an alternative approach to obtain a sparse $\mathbf{W}$ by directly enforcing sparse pivot predictors $\mathbf{w}_l$ through L1-regularization (Tsuruoka et al., 2009), but didn't pursue this strategy due to unstable results. Since SGD is sensitive to feature scaling the projection $\theta\mathbf{x}$ 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 α such that $|D_{\mathcal{S}}|^{-1} \sum_{\mathbf{x} \in D_{\mathcal{S}}} \|\alpha\theta\mathbf{x}\| = 1$.

We use Google Translate as word translation oracle, which returns a single translation for each query word.[5] Though such a context free translation is suboptimum we do not sanitize the returned words to demonstrate the robustness of CL-SCL with respect to translation noise. To ensure the reproducibility of our results we cache all queries to the translation oracle.

[1] http://www.webis.de/research/corpora/webis-cls-10/

[2] http://mecab.sourceforge.net

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

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

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

$\mathcal{T}$	Category	Unlabeled data		Upper Bound		CL-MT			CL-SCL		
		$\lvert D_{\mathcal{S},u}\rvert$	$\lvert D_{\mathcal{T},u}\rvert$	μ	σ	μ	σ	Δ	μ	σ	Δ
German	books	50,000	50,000	83.79	(±0.20)	**79.68**	(±0.13)	4.11	79.50	(±0.33)	4.29
	dvd	30,000	50,000	81.78	(±0.27)	**77.92**	(±0.25)	3.86	76.92	(±0.07)	4.86
	music	25,000	50,000	82.80	(±0.13)	77.22	(±0.23)	5.58	**77.79**	(±0.02)	5.00
French	books	50,000	32,000	83.92	(±0.14)	**80.76**	(±0.34)	3.16	78.49	(±0.03)	5.43
	dvd	30,000	9,000	83.40	(±0.28)	**78.83**	(±0.19)	4.57	78.80	(±0.01)	4.60
	music	25,000	16,000	86.09	(±0.13)	75.78	(±0.65)	10.31	**77.92**	(±0.03)	8.17
Japanese	books	50,000	50,000	79.39	(±0.27)	70.22	(±0.27)	9.17	**73.09**	(±0.07)	6.30
	dvd	30,000	50,000	81.56	(±0.28)	**71.30**	(±0.28)	10.26	71.07	(±0.02)	10.49
	music	25,000	50,000	82.33	(±0.13)	72.02	(±0.29)	10.31	**75.11**	(±0.06)	7.22

Table 1: Cross-language sentiment classification results. For each task, the number of unlabeled documents from $\mathcal{S}$ and $\mathcal{T}$ is given. 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. Δ gives the difference in accuracy to the upper bound. CL-SCL uses $m = 450$, $k = 100$, and $\phi = 30$.

5.3 Upper Bound and Baseline

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

We chose a machine translation baseline to compare CL-SCL to another cross-language method. Statistical machine translation technology offers a straightforward solution to the problem of 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). Our baseline CL-MT works as follows: (1) learn a linear classifier on the training data, and (2) translate the test documents into the source language,[6] (3) predict the sentiment polarity of the translated test documents. Note that the baseline CL-MT does not make use of unlabeled documents.

[6] Again we use Google Translate.

5.4 Performance Results and Sensitivity

Table 1 contrasts the classification performance of CL-SCL with the upper bound and with the baseline. Observe that the upper bound does not exhibit a great variability across the three languages. The average accuracy is about 82%, which is consistent with prior work on monolingual sentiment analysis (Pang et al., 2002; Blitzer et al., 2007). The performance of CL-MT, however, differs considerably between the two European languages and Japanese: for Japanese, the average difference between the upper bound and CL-MT (9.9%) is about twice as much as for German and French (5.3%). This difference can be explained by the fact that machine translation works better for European than for Asian languages such as Japanese.

Recall that CL-SCL receives three hyperparameters as input: the number of pivots m, the dimensionality of the cross-lingual representation k,

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

Table 2: Semantic and pragmatic correlations identified for the two pivots {beautiful$_{\mathcal{S}}$, schön$_{\mathcal{T}}$} and {boring$_{\mathcal{S}}$, langweilig$_{\mathcal{T}}$} in English and German book reviews.

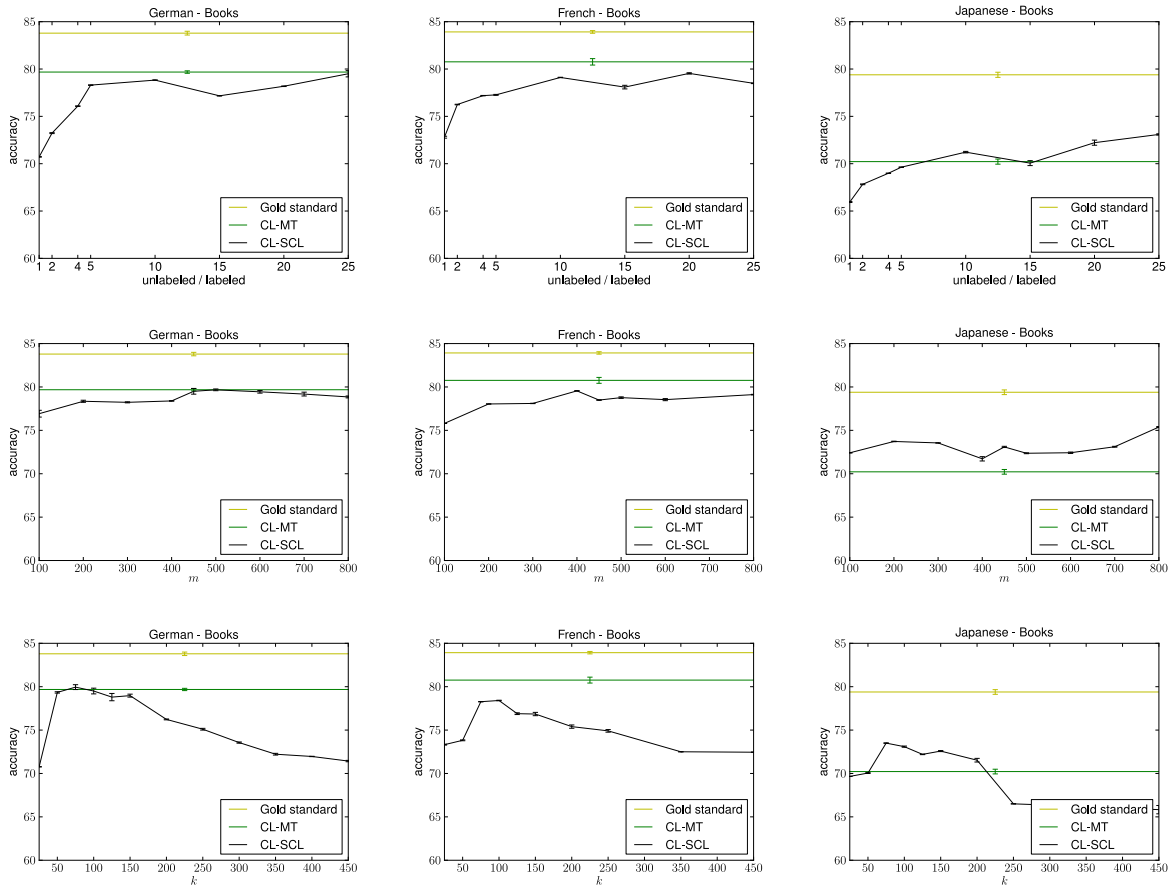

Figure 2: Influence of unlabeled data and hyperparameters on the performance of CL-SCL. The rows show the performance of CL-SCL as a function of (1) the ratio between labeled and unlabeled documents, (2) the number of pivots m, and (3) the dimensionality of the cross-lingual representation k.

and the minimum support ϕ of a pivot in $D_{\mathcal{S},u}$ and $D_{\mathcal{T},u}$. For comparison purposes we use fixed values of $m = 450$, $k = 100$, and $\phi = 30$. The results show the competitiveness of CL-SCL compared to CL-MT. Although CL-MT outperforms CL-SCL on most tasks for German and French, the difference in accuracy can be considered as small (<1%); merely for French book and music reviews the difference is about 2%. For Japanese, however, CL-SCL outperforms CL-MT on most tasks with a difference in accuracy of about 3%. The results indicate that if the difference between the upper bound and CL-MT is large, CL-SCL can circumvent the loss in accuracy. Experiments with language-specific settings revealed that for Japanese a smaller number of pivots ($150<m<250$) performs significantly better. Thus, the reported results for Japanese can be considered as pessimistic.

Primarily responsible for the effectiveness of CL-SCL is its task specificity, i.e., the ways in which context contributes to meaning (pragmatics). Due to the use of task-specific, unlabeled data, relevant characteristics are captured by the pivot classifiers. Table 2 exemplifies this with two pivots for German book reviews. The rows of the table show those words which have the highest correlation with the pivots $\{\text{beautiful}_{\mathcal{S}}, \text{schön}_{\mathcal{T}}\}$ and $\{\text{boring}_{\mathcal{S}}, \text{langweilig}_{\mathcal{T}}\}$. We 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 connection that authors of book reviews tend to use the word “beautiful” to refer to illustrations or poetry. While the first type of word correlations can be obtained by methods that operate on parallel corpora, the second type of correlation requires an understanding of the task-specific language use.

In the following we discuss the sensitivity of each hyperparameter in isolation while keeping

the others fixed at $m = 450$, $k = 100$, and $\phi = 30$. The experiments are illustrated in Figure 2.

Unlabeled Data The first row of Figure 2 shows the performance of CL-SCL as a function of the ratio of labeled and unlabeled documents. 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 1. As expected, an increase in unlabeled documents results in an improved performance, however, we observe a saturation at a ratio of 10 across all nine 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 of pivots captures a significant amount of the correspondence between $\mathcal{S}$ and $\mathcal{T}$.

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. 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 $75{<}k{<}150$ works equally well across all tasks.

6 Conclusion

The paper introduces a novel approach to cross-language text classification, called cross-language structural correspondence learning. The approach uses unlabeled documents along with a word translation oracle to automatically induce task-specific, cross-lingual correspondences. Our contributions include the adaptation of SCL for the problem of cross-language text classification and a well-founded empirical analysis. The analysis covers performance and robustness issues in the context of cross-language sentiment classification with English as source language and German, French, and Japanese as target languages. The results show that CL-SCL is competitive with state-of-the-art machine translation technology while requiring fewer resources.

Future work includes the extension of CL-SCL towards a general approach for cross-lingual adaptation of natural language processing technology.

References

Rie-K. Ando and Tong Zhang. 2005a. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853.

Rie-K. Ando and Tong Zhang. 2005b. A high-performance semi-supervised learning method for text chunking. In *Proceedings of ACL-05*, pages 1–9, Ann Arbor.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of ICWSM-08*, pages 19–26, Seattle.

Nuria Bel, Cornelis H. A. Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *Proceedings of ECDL-03*, pages 126–139, Trondheim.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of EMNLP-06*, pages 120–128, Sydney.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL-07*, pages 440–447, Prague.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL-07*, pages 256–263, Prague.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on CrossLanguage Text and Speech Retrieval*.

Jenny-R. Finkel and Christopher-D. Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of HLT/NAACL-09*, pages 602–610, Boulder.

Blaž Fortuna and John Shawe-Taylor. 2005. The use of machine translation tools for cross-lingual text mining. In *Proceedings of the ICML Workshop on Learning with Multiple Views*.

Alfio Gliozzo and Carlo Strapparava. 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*.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of ACL-06*, pages 553–560, Sydney.

Kanayama Hiroshi, Nasukawa Tetsuya, and Watanabe Hideo. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of COLING-04*, pages 494–500, Geneva.

Jing Jiang and Chengxiang Zhai. 2007. A two-stage approach to domain adaptation for statistical classifiers. In *Proceedings of CIKM-07*, pages 401–410, Lisbon.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *Proceedings of SIGIR-02*, pages 175–182, Tampere.

Yaoyong Li and John S. Taylor. 2007. Advanced learning algorithms for cross-language patent retrieval and classification. *Inf. Process. Manage.*, 43(5):1183–1199.

Xiao Ling, Gui-R. Xue, Wenyuan Dai, Yun Jiang, Qiang Yang, and Yong Yu. 2008. Can chinese web pages be classified with english data source? In *Proceedings of WWW-08*, pages 969–978, Beijing.

Douglas W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of AMTA-98*, pages 472–483, Langhorne.

J. Scott Olsson, Douglas W. Oard, and Jan Hajič. 2005. Cross-language text classification. In *Proceedings of SIGIR-05*, pages 645–646, Salvador.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02*, pages 79–86, Philadelphia.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A wikipedia-based multilingual retrieval model. In *Proceedings of ECIR-08*, pages 522–530, Glasgow.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2007. Learning visual representations using images with captions. In *Proceedings of CVPR-07*, pages 1–8, Minneapolis.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An em based training algorithm for cross-language text categorization. In *Proceedings of WI-05*, pages 529–535, Compiègne.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. 2007. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of ICML-07*, pages 807–814, Corvalis.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of ACL/AFNLP-09*, pages 477–485, Singapore.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL/AFNLP-09*, pages 235–243, Singapore.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of ICML-04*, pages 116–124, Banff.