Cross-lingual Text Classification via Model Translation with Limited Dictionaries

Ruochen Xu
Language Technologies Inst.
Carnegie Mellon University
Pittsburgh, PA 15213, USA
ruochenx@cs.cmu.edu

Yiming Yang
Language Technologies Inst.
Carnegie Mellon University
Pittsburgh, PA 15213, USA
yiming@cs.cmu.edu

Hanxiao Liu
Language Technologies Inst.
Carnegie Mellon University
Pittsburgh, PA 15213, USA
hanxiaol@cs.cmu.edu

Andrew Hsi
Language Technologies Inst.
Carnegie Mellon University
Pittsburgh, PA 15213, USA
ahsi@cs.cmu.edu

ABSTRACT

Cross-lingual text classification (CLTC) refers to the task of classifying documents in different languages into the same taxonomy of categories. An open challenge in CLTC is to classify documents for the languages where labeled training data are not available. Existing approaches rely on the availability of either high-quality machine translation of documents (to the languages where massively training data are available), or rich bilingual dictionaries for effective translation of trained classification models (to the languages where labeled training data are lacking). This paper studies the CLTC challenge under the assumption that neither condition is met. That is, we focus on the problem of translating classification models with highly incomplete bilingual dictionaries. Specifically, we propose two new approaches that combines unsupervised word embedding in different languages, supervised mapping of embedded words across languages, and probabilistic translation of classification models. The approaches show significant performance improvement in CLTC on a benchmark corpus of Reuters news stories (RCV1/RCV2) in English, Spanish, German, French and Chinese and an internal dataset in Uzbek, compared to representative baseline methods using conventional bilingual dictionaries or highly incomplete ones.

Keywords

Multilingual Text Data; Cross-lingual Text Classification; Transfer Learning

1. INTRODUCTION

The massive amount of multilingual documents on the World Wide Web makes the cross-lingual text categorization (CLTC) problem increasingly important, whose solutions aim to provide organizational views of the data. Typically, CLTC refers to the task of classifying documents in different languages using the same taxonomy of predefined categories. The Reuters News Agency, for example, has been using the same taxonomy of subject topics to index International news stories in different languages. Automated classification of multilingual documents is obviously desirable for both cost saving and classification consistency.

The CLTC problem would be relatively easy to solve if we had a sufficient amount of labeled training data for each language because most machine learning techniques for text classification have the flexibility to be applied to any language. However, for many languages in the real world, a large quantity of human-labeled documents for training classifiers is often hard to obtain. Thus a natural solution is to train classifiers in a label-rich language and then apply the trained classifiers to documents in label-poor languages. For convenience let us denote the language that provides labeled documents for training classifiers as the source language, and the other languages that provide unlabeled test documents as the target languages. How to successfully apply the trained classifiers in the source language to documents in different target languages is the key question for research.

Existing CLTC methods differ in how to make the classification across languages. Bel et al. [2] presented an early effort where they translated the target-language documents to the source language using an comprehensive bilingual dictionary, and then applied the classifiers in the source language to the translated documents. To reduce the computational cost, they only translated the topically important terms in those documents. Similarly, Ling et al. [13] also translated target-language documents (Chinese web pages) to a source language (English), and predicted their labels based the labels of the English documents which are similar to the translated versions of the Chinese documents.

Rigutini et al. [21] translated training documents from a source language to a target language instead, and applied an Expectation Maximization (EM) algorithm to leverage unlabeled documents in the target language in addition. In the E-step the unknown labels were guessed for the target-language documents, and in the M-step the classifier parameters were updated based on both the (translated) training documents with true labels and the target-language documents with guessed labels.

Wan [26] used machine translation (MT) systems to perform English-to-Chinese and Chinese-to-English translation of each document in a collection of labeled English and unlabeled Chinese documents. The original document (before translation) and its

CIKM'16 , October 24 - 28, 2016, Indianapolis, IN, USA

ISBN 978-1-4503-4073-1/16/10. . . $15.00

DOI: http://dx.doi.org/10.1145/2983323.2983732

translated version were called the *two views* of the same documents. The two views of all the documents enabled a co-training algorithm to train and re-train classifiers both in English and Chinese alternately and iteratively. That is, it started with the labeled portion of the documents as the initial training set, and then added more and more classifier-assigned labels to the unlabeled portion of the documents for retraining. This process resulted in improved classifiers in both languages while human-labeled documents were available only in one language. With the help of machine translation, some recent works also solved CLTC via multi-view learning methods, including majority voting[1], multi-view co-regularization[8] and representation learning[9].

Instead of translating documents as in the above approaches, Shi et al. [22] tried to translate classification models across languages. The source-language model of each category consisted of a bag of weighted terms, where the term weights were the learned model parameters based on labeled data. Then each term in the model was translated to the target language based on a comprehensive bilingual dictionary. To handle ambiguities (one-to-many mapping) in term translation, an EM algorithm was used to obtain the cross-lingual translation probabilities.

While the relevant literature has provided valuable insights about how to tackle the CLTC problem, existing methods have an implicit or explicit assumption in common, i.e., the availability of rich cross-lingual knowledge resources for each language pair of interest. By rich knowledge resources here we mean comprehensive bilingual dictionaries and MT systems for quality-translation of documents or classification models in the domains of interest. Such an assumption would significantly limit the generalization or applicability of those methods to a broad range of low-resource languages. In fact, except the dominating or most common languages (like English, French, Spanish, etc.), the majority of languages in the real world often do not have large quantities of comprehensive cross-lingual dictionaries or high-quality MT systems to support CLTC in every possible domain of interest. This fact makes the CLTC challenge wildly open, i.e., we must solve the problem without relying on the availability of rich cross-lingual knowledge resources. How do we get there? Existing research in CLTC has not answered this question.

This paper focuses on the open challenge of low-resource CLTC, especially under the condition where the bilingual dictionaries are highly incomplete and very small in size. We further narrow down our focus on translating classification models across languages instead of translating documents, as the former is computationally much more efficient than the latter (when the document collections are very large), and often the solutions of the former can be easily generalized to the latter in principle.

We propose a new approach that combines the strengths of unsupervised word embedding in multilingual documents, supervised mapping of embedded words across languages for bilingual dictionary extension, and probabilistic aggregation of word alignments for translating classification models. Our idea is partly inspired by the recent research in distributional word embedding with multilingual documents or parallel text[3, 5, 24] and the application to a variety of tasks, including machine translation [27], cross-lingual word alignment [11, 17, 25, 23], dependency parsing[4, 7, 6, 23] and more.

Although CLTC has become the standard benchmark to evaluate multilingual word embeddings[10, 20], the purpose is not to induce a state-of-art cross-lingual classifier, but rather to examine the informativeness of the induced representations. In our experiments, we find that the simple way to train cross-lingual classifier in [10, 20] is sub-optimal. In this work, however, we focus on making accurate CLTC predictions, especially under low resource conditions.

The rest of the paper is organized as follows. Section 2 describes our word-embedding based methods for bilingual dictionary extension, as a component of our approach. Section 3 describes our method for probabilistic translation of classification models. Section 4 outlines our controlled experiments for evaluating the proposed approach in comparison with a representative baseline, on a multilingual benchmark data set (RCV2) under simulated low-resource conditions and an internal dataset in a real low-resource language. Section 5 analyzes the experiment results and section 6 concludes with the major findings in this paper.

2. CROSS-LINGUAL DICTIONARY EXTENSION

We explore two new approaches to statistical extension of bilingual dictionaries for CLTC, conditioned on the availability of a small-sized (incomplete) dictionary per language pair. Both methods combine strengths of unsupervised word embedding in each language and supervised or semi-supervised mapping of words across languages. The two methods differ in how to establish the mapping. The first method, regularized linear regression or Ridge regression (RidgeReg), uses supervised learning to obtain a set of ridge regression models for cross-lingual word mapping, where the true translation pairs of words in the initial dictionary are treated as labeled training pairs. The second is a transductive learning method that jointly leverages both labeled and unlabeled word pairs across two languages in the optimization of the mapping. We call it the *Transductive Label Propagation (TransLP)* approach. Neither methods have been studied for dictionary extension in CLTC, to our knowledge. In addition to the two methods we propose, we also include a representative state-of-art proposed by Faruqui and Chris[3], which we call the *Canonical Correlation Analysis(CCA)* method, and a simple and intuitive baseline for comparison, which we call the *k Nearest Neighbor (kNN)* method.

2.1 Regularized Linear Regression (RidgeReg)

Word embedding has been a hot topic in recent machine learning and was found effective in capturing semantic similarities among words when trained on large document collections. It discovers a vector representation of each word based on its co-occurrence patterns with other words, and enables inference about some semantic relations via simple operations. For example, *vector(England) - vector(London)* would be similar to *vector(China) - vector(Beijing)* [19]. More interestingly, Mikolov et al. [17] showed that such linear relations exist in different languages, and the relations can be easily translated across languages using multivariate linear regression.

Intrigued by this line of work we propose to extend an existing (small-sized) bilingual dictionary via cross-lingual regression over embedded words. Denote by $\mathcal{D} = \{(x_i, z_i)\}_{i=1}^{n}$ the given bilingual dictionary, where $x_i \in \mathbb{R}^{d_1}$ is the vector representation of word i in the source language and $z_i \in \mathbb{R}^{d_2}$ is the vector representation of word j in the target language with equivalent meaning. Using the word pairs in the dictionary as the labeled training set, we can learn a transformation matrix with the following objective function:

$$\min_{W \in \mathbb{R}^{d_2 \times d_1}} \sum_{i=1}^{n} \|Wx_i - z_i\|^2 + \lambda \|W\|_F^2 \qquad (1)$$

Here W is the unknown matrix we want to optimize; the first term

in the function is the training-set loss, and the second is the regularization term, to avoid overfitting on the training data. Notice that optimizing W row-by-row is equivalent to solving a series of ridge regression problems.

The problem in (1) has a closed-form solution given by

$$W^* = ZX^T(XX^T + \lambda I)^{-1} \tag{2}$$

where $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d_1 \times n}$ and $Z = [z_1, z_2, ..., z_n] \in \mathbb{R}^{d_2 \times n}$.

Once W is learned based on the training set, we can use it to map any x which may be unseen in $\mathcal{D}$ to vector $z = Wx$ in the target space. Then we can find the target-language words as the translations of word x if those words are among the k nearest neighbors of vector z. We treat k as a hyper parameter of this method, which could be tuned via cross-validation; in our empirical evaluation we find performance is not sensitive to the choice of k and set it to 10 for all experiments. We use the cosine similarity as the nearness measure among vectors.

2.2 Transductive Label Propagation (TransLP)

Although ridge regression is a natural choice for establishing a linear mapping of embedded words across languages, it does not explore the power of non-linear transformation. Also it only leverages the labeled data (the true translation pairs of words in the given dictionary), but not the vastly available unlabeled words in both the source language and the target language. To broaden the scope of our investigation, we propose a transductive label propagation (TransLP) approach for a non-linear crosslingual mapping and for utilizing both labeled and unlabeled data during training.

TransLP is a semi-supervised learning framework that has been developed recently for bipartite link prediction based on multi-source relations [14]. The key idea is to use the graph product operations (such as the Kronecker product) to combine relational information in multi-source graphs, and then to propagate the label information in the observed (labeled) links to the unknown (unlabeled) links over the product graph. Adapting this idea to cross-lingual mapping of embedded words, we want to propagate the labels of the known translation pairs of words (in the provided dictionary) to the unknown pairs based on word-word similarities within both the source language and the target language.

Denote by $G \in \mathbb{R}^{|V| \times |V|}$ and by $H \in \mathbb{R}^{|V'| \times |V'|}$ the word similarity graphs within the source and target languages, respectively, where $G_{ii'}$ encodes the similarity between word i and word i' in the source language, and $H_{jj'}$ encodes the similarity between word j and word j' in the target language. Specifically, we construct these graphs by computing the pairwise cosine similarities for the embedded words within each language . We further define the induced similarity between the cross-lingual links (i, j) and (i', j') as $G_{ii'}H_{jj'}$, which means that the two links should have similar labels (as word translation pairs or not) if the embeddings of words i and i' are similar in the source language and if embeddings of words j and j' are similar in the target language. An illustration of TransLP with some toy data is shown in figure 1

Denote by F_{ij} the system-predicted score for cross-lingual link (i, j). Intuitions above can be encoded in the following Gaussian random field prior over $F \in \mathbb{R}^{|V| \times |V'|}$

$$vec(F) \sim \mathcal{N}(0, G \otimes H) \tag{3}$$

where vec is the vectorization operator that concatenates the columns of a matrix into a single vector, $G \otimes H$ stands for the Kronecker product of G and H.

Our optimization objective is defined as

$$\min_{F \in \mathbb{R}^{|V| \times |V'|}} \quad \sum_{i,j} \ell\left(F_{ij}, \mathbf{1}_{(x_i, z_j) \in \mathcal{D}}\right) + \frac{\gamma}{2} r(F) \tag{4}$$

where regularization $r(F)$ corresponds to the negative likelihood of the Gaussian random field prior defined by (3)

$$r(F) = vec(F)^\top (G \otimes H)^\dagger vec(F) \propto -\log p(F \mid G, H) + const \tag{5}$$

where † stands for matrix pseudoinverse.

One may choose any loss function ℓ in (4), such as squared error $\ell(x, y) := (x - y)^2$, the indicator function $\mathbf{1}_{\{\cdot\}}$ equals 1 if $(x_i, z_j) \in \mathcal{D}$ and equals zero otherwise. The first term in (4) encourages our predictions to fit the observed labels, and the second term encourages the predicted values to have a smooth propagation with respect to the similarities among cross-lingual word pairs.

Optimizing (4) would be extremely expensive when $|V|$ and $|V'|$ are large. To speedup, we propose to reduce the computational complexity via low-rank approximations. More specifically, we approximate G and H with their leading eigenvectors $U \in \mathbb{R}^{|V| \times k_1}$ and $V \in \mathbb{R}^{|V'| \times k_2}$, and restrict matrix F within the linear span of those eigenvectors

$$G = \sum_{i=1}^{k_1} \lambda_i v_i v_i^\top \tag{6}$$

$$H = \sum_{j=1}^{k_2} \mu_j u_j u_j^\top \tag{7}$$

$$F = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \alpha_{ij} v_i u_j^\top \tag{8}$$

For sparse graphs, both U and V can be obtained via power iteration algorithm in linear complexity over $|V|$ and $|V'|$. The regularization term in (4) can be simplified as

$$\begin{aligned}
r(F) =& vec(F)^\top (G \otimes H)^\dagger vec(F) && (9)\\
=& vec(F)^\top (G^\dagger \otimes H^\dagger) vec(F) && (10)\\
=& \langle F, G^\dagger F H^\dagger \rangle && (11)\\
=& \Big\langle F, \sum_{i=1}^{k_1} \lambda_i^\dagger v_i v_i^\top \sum_{i'=1}^{k_1} \sum_{j'=1}^{k_2} \alpha_{i'j'} v_{i'} u_{j'}^\top \sum_{j=1}^{k_2} \mu_j^\dagger u_j u_j^\top \Big\rangle && (12)\\
=& \Big\langle \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \alpha_{ij} v_i u_j^\top, \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \alpha_{ij} \lambda_i^\dagger \mu_j^\dagger v_i u_j^\top \Big\rangle && (13)\\
=& \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \alpha_{ij}^2 \lambda_i^\dagger \mu_j^\dagger && (14)
\end{aligned}$$

We then solve the following optimization problem with a substantially reduced number of model parameters

$$\begin{aligned}
\min_{\{\alpha_{ij}\}_{i=1,j=1}^{k_1,k_2}} \quad & \sum_{i,j} \ell\left[\left(\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \alpha_{ij} v_i u_j^\top\right)_{ij}, \mathbf{1}_{(x_i, z_j) \in D}\right] \\
& + \frac{\gamma}{2} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \alpha_{ij}^2 \lambda_i^\dagger \mu_j^\dagger
\end{aligned} \tag{15}$$

It is not hard to verify that optimization problem (15) is convex over the α_{ij}'s. Since typically $k_1, k_2 \ll \min\{|V|, |V'|\}$, each gradient update for the above optimization only takes $O(|V| +$

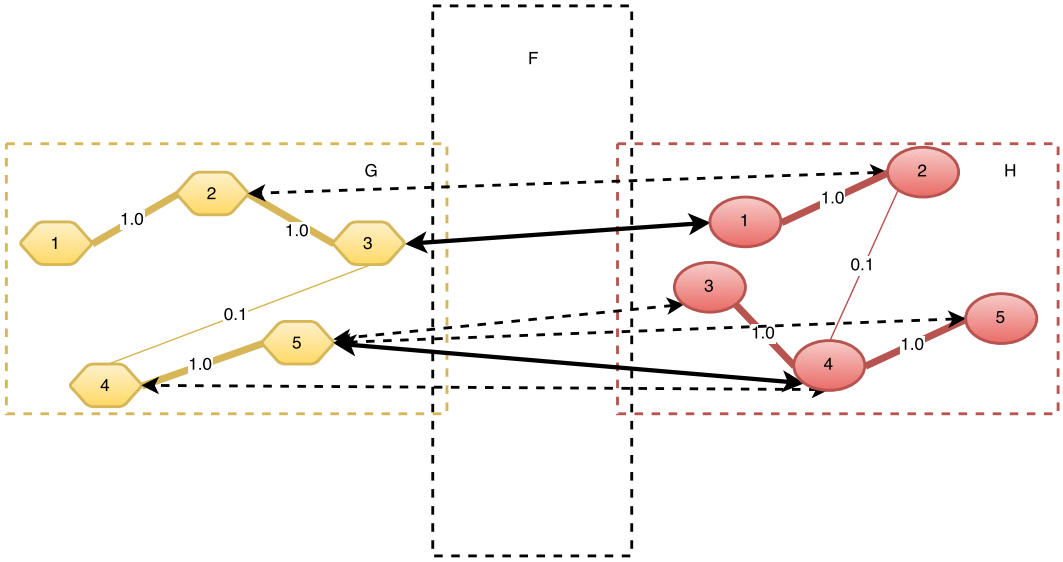

Figure 1: Illustration of TransLP with toy examples. Nodes in different color and shape represent words in different languages. G and H contain edges encoded by word-to-word similarity. Solid lines with double arrows are observed bilingual dictionary pairs and dashed lines with double arrows are predicted bilingual pairs.

$|V'|$) flops. We empirically find it sufficient to set $k_1 \leq 500, k_2 \leq 500$ for good performance in practice.

2.3 Baselines

2.3.1 *Canonical Correlation Analysis(CCA)*

We follow the work of Faruqui and Chris[3] as comparison for our proposed methods with state-of-art approach of cross-lingual word embeddings. They used canonical correlation analysis(CCA) for incorporating multilingual evidence into vectors generated monolingually. Given mapped monolingual word vectors X and Z as defined in (2), CCA first seeks v and w such that:

$$v, u = \text{argmax}_{v \in \mathbb{R}^{d_1}, u \in \mathbb{R}^{d_2}} corr(v^T X, w^T Z) \tag{16}$$

$v^T X, w^T Z$ are called the first canonical variate pair. We further seeks vectors maximizing the same correlation but subject to the constraint that they are to be uncorrelated with the first canonical variate pair. This process may continue to d times, where $d = \min(d_1, d_2)$. The resulting matrix $V \in \mathbb{R}^{d_1 \times d}$ and $W \in \mathbb{R}^{d_2 \times d}$ are used to map any source word vector x and target word vector z, which may not appear in the bilingual dictionary, to a unified space: $x^\star = V^T x, z^\star = W^T z$. We apply the same procedure to find translations in the unified vector space as described in 2.1.

2.3.2 *K-Nearest Neighbor (kNN)*

As an intuitive and simple baseline for comparison, the kNN approach for bilingual dictionary extension is defined as the following. For each word pair (w', w) in the initial bilingual dictionary where w' is a word in the source language and w is a true translation of w' in the target language, we extend the dictionary by adding the k words most similar to w in the target language as the valid translations of word w'. Symmetrically, we do such kNN extension for word w' in the source language as well. Notice that each word has a vector representation obtained by applying word embedding to each language, and that the similarity between each word pair in the language is measured by the cosine of the corresponding vectors.

3. CLASSIFICATION MODEL TRANSLATION

3.1 Cross-lingual Naive Bayes(CLNB)

Given a bilingual dictionary (extended using the methods in the above section), we want to translate the classification models trained on the labeled documents in the source language to the target language. We use a standard multinomial Naïve Bayes (NB) as the classification method [15] because the probabilistic model parameters allow easy translation of NB models in a probabilistic manner[1].

Given document d in the target language, the conditional probability of d being generated from category c is given by:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{17}$$

Since $P(d)$ is independent of c, the denominator of (17) can be ignored when predicting category label $\hat{y}$ for document d as

$$\hat{y} = \arg\max_{c \in C} P(c)P(d|c) \tag{18}$$

where C is the candidate set of category labels, and $P(d|c)$ is the probability of document d conditioned on category c. The latter is proportional to the product of word probabilities under the independence assumption:

$$P(d|c) \propto \prod_{w \in d} P(w|c) \tag{19}$$

For engaging with cross-lingual translation probabilities, we specify word probability $P(w|c)$ to be decomposed as:

$$P(w|c) = \sum_{w'} P(w', w|c) = \sum_{w'} P(w'|c)P(w|w', c) \tag{20}$$

[1]We have also examined other types of classifiers including Support Vector Machines (SVM), and found that the associated model translation is either more complicated or less effective, or both. Details on this are beyond the scope of this paper.

where word w' is any word in the source language, $P(w'|c)$ is the probability of source word conditioned on category c, and $P(w|w', c)$ is the translational probability form w' to w conditioned on category c.

Given a training set of labeled documents in the source language with the vocabulary size of V', conditional probability $P(w'|c)$ is typically estimated with Laplace smoothing as:

$$\hat{P}(w'|c) = \frac{T_{c,w'} + 1}{(\sum_{v' \in V'} T_{c,v'}) + |V'|} \quad (21)$$

where $T_{c,w'}$ be the number of occurrences of word w' in the training documents from category c. As for category prior, we just use the Maximum Likelihood Estimate (MLE) of $\hat{P}(c) = \frac{N_c}{N}$ where N is the size of the labeled training set and N_c is the number of labeled documents in category c. We assume the category priors are the same in both the source and the target languages.

Now the missing part we need for completing formula (20) is $P(w|w', c)$. Denoting by $\mathcal{D}$ the given bilingual dictionary, we set $P(w|w', c) = 0$ if pair $(w', w) \notin \mathcal{D}$; otherwise, we estimate it using cross-lingual word similarities with normalization as:

$$P(w|w', c) \approx P(w|w') \approx \frac{sim(w, w')}{\sum_{v \in D(w')} sim(v, w')} \quad (22)$$

where $D(w')$ is the set of target-language words as the translations of source word w' in the dictionary; $sim(w, w')$ is the similarity score given by dictionary extension methods for target word w and source word w'. For example, in RidgeReg we have

$$sim(w, w') = cos(Wx', x)$$

where x' and x are vector representation for w' and w. Similarly, for CCA we have

$$sim(w, w') = cos(V^T x', W^T x)$$

In TransLP we have

$$sim(w, w') = F_{ij}$$

where i and j are indices for w and w'. And in kNN we have

$$sim(w, w') = \begin{cases} 1, & \text{if } (w, w') \in \mathcal{D} \\ 0, & \text{otherwise} \end{cases}$$

Our model translation method is computationally very efficient. At testing time, $P(w|c)$ is first computed according to equation (20) and stored. Then the classifier makes predictions following equations (18) and (19) in the same way as in standard monolingual Naïve Bayes classifiers. In practice, it takes less than a second to run our CLTC method over a test collection of a few thousand documents.

A potential weakness of our method, on the other hand, is in its approximation of $P(w|w', c) \approx P(w|w')$. That is, the word embedding components we used and the cross-lingual mapping of embedded words are not category-sensitive. We leave the category-sensitive enhancement of our approach to future research.

3.2 Baselines

3.2.1 *Cross-Lingual Mixture Model(CLMM)*

The method by Shi et al.[22] also estimates $P(w|w', c)$ to translate their classification model from source language to target language. They exploited the readily available unlabeled data in the target language via semi-supervised learning. To summarize, they optimize $\theta = P(w|w', c)$ to maximize the following log-likelihood:

$$l(\theta) = \sum_{d \in D_u} log \sum_c p(c) \sum_{d' \in D(d)} \prod_{w' \in d'} P(w|w', c) p(w'|c) \quad (23)$$

where $p(w'|c)$ and $p(c)$ are learned from training data in source language and viewed as fix parameters; $d' \in D(d)$ represents all possible document d' translated from d according to dictionary $\mathcal{D}$. Note that the model assumes the availability of certain amount of unlabeled documents in target language(i.e. D_u). Those documents are further required to belong to the same taxonomy of data in source language. Those assumptions may not hold in the low-resource scenario. On the other hand, our proposed model is capable of utilizing more accessible general-purpose monolingual corpus(e.g. Wikipedia) to propagate existing bilingual dictionary.

3.2.2 *Dimension Reduction(DR)*

The method was used to show the effectiveness and informativeness of cross-lingual word embeddings [10, 20]. Suppose we have $Universal\_Vec$ as the learned word representation for both source language and target language. $Universal\_Vec(w)$ returns the vector for any word w in either languages. Instead of using sparse bag-of-word feature for documents, we represent each document d in both source and target language as

$$\sum_{w \in d} tfidf(w) \cdot Universal\_Vec(w)$$

Using the unified representation, model trained on source data could be directly applied on target data. In our experiment, we implemented this method with the output vector from $RidgeReg$. The classifier was chosen to be an averaged perceptron[2] as used in [10, 20]. We denote this baseline as $DR.RidgeReg$ in the following evaluation.

4. EMPIRICAL EVALUATION DESIGN

Our experiments include the evaluation of the proposed methods in bilingual dictionary extension (as a sub-task), and in cross-lingual text classification as the end-to-end evaluation. We fixed English as source language. Since it is hard to obtain a large labeled dataset in real low-resource language, we used Spanish, French, German and Chinese available in RCV2[12] to simulate low-resource condition. For completeness, we also include another smaller internal dataset in Uzbek, a real low-resource language, to prove the effectiveness of our methods.

4.1 Evaluation for bilingual dictionary extension

We obtained online bilingual dictionaries from English to Spanish, French and German via *MyMemory* [3] and from English to Chinese via *CC-CEDICT* [4]. The English-Uzbek dictionary was given in the internal dataset. The dictionary sizes are measured using the number of word translation pairs, as summarized in Table 1; the branching factor means the average number of translated words in the target language per source word.

For word embedding in English, we directly used the pre-trained vectors on a Google News dataset[5]. For the other languages, we

[2]We also tried with SVM as a stronger classifier, but it gave similar performance as averaged perceptron.
[3]http://mymemory.translated.net/
[4]https://www.mdbg.net/chindict/chindict.php?page=cedict
[5]https://code.google.com/p/word2vec/

Table 1: Statistics of the bilingual dictionaries

Target Language	Size	Branching Factor
Spanish	11518	2.21
French	9901	2.05
German	8856	2.00
Chinese	8185	2.31
Uzbek	9066	2.35

applied the Continuous Bag-of-Words Model [18, 16] to an unlabeled corpus in each language. Specifically, for Spanish, French, German and Chinese, we used subsets of multilingual Wikipedia pages[6]; for Uzbek, monolingual text is harvested from the web, which includes news text, blogs, discussion forums, Twitter and reference materials like Wikipedia. Table 2 summarizes sizes of these monolingual corpora.

Table 2: The sizes of the monolingual corpora

Language	Tokens	Vocabulary Size
English	100B	3M
Spanish	412M	665K
French	488M	754K
German	619M	1505K
Chinese	123M	723K
Uzbek	52M	510K

To simulate the low-resource conditions we sub-sampled 1%, 10%, 25%, 50%, 75% and 100% of the translation pairs in each bilingual dictionary. Specifically, for each fixed percentage we randomly sub-sampled 10 times from the pool, and averaged the performance scores of each method over these ten samples. Each sample was further split into subsets of 50% for training, 25% for validation set (for tuning parameters) and 25% for testing. For all runs, the λ of RidgeReg was set to 1 and γ of TransLP was set to 10^{-6}.

We evaluated the performance on each test set using a ranking metric. That is, we treated each target-language word in the test set as query, and its true translation in the source language as the relevant items. For each true translation pair, we randomly sampled 100 words in the source language as the irrelevant items. The union of all the relevant and irrelevant items of each query form the candidate set for the query. Each candidate was scored by one of the dictionary extension methods (RidgeReg, TransLP, CCA or kNN). By sorting the scores with respect to each query we obtained a ranked list per query. We then evaluated the ranked lists using the mean average precision (MAP), which is conventional in evaluation of retrieval systems. The MAP scores range from 0 to 1; higher MAP means the better performance.

4.2 Evaluation of the CLTC performance

4.2.1 *RCV1/RCV2 Dataset*

We used the Reuters RCV1/RCV2 [12] benchmark corpora for this part of the evaluation. RCV1 contains a large number of English news stories and each document belongs to at least one topical category. RCV2 includes news stories in several languages (including the four "low-resource" languages in our study) with topic labels in the same taxonomy. However, the document collections

[6]https://sites.google.com/site/rmyeid/projects/polyglot#TOC-Download-Wikipedia-Text-Dumps

Table 4: Statistics of RCV1 and RCV2. Size refers to the number of documents. Topic Categories containing less than 5 documents are discarded

	Language	Size	Num. of Categories
RCV1	English	23149	98
RCV2	Spanish	18655	64
	French	20000	70
	German	20000	71
	Chinese	28964	61

Table 5: Statistics of Uzbek dataset. Size refers to the number of documents.

Genre	Language	Size
News Article	English	1218
Discussion Forum	English	501
News Article	Uzbek	49893
Discussion Forum	Uzbek	12898

are not parallel, i.e., they are not translations of each other. We used a subset of the English, French and German documents and all the available documents in Spanish and Chinese in our experiments. The statistics of these datasets are shown in table 4. All the topic categories in RCV2 are covered by the set of categories contained in RCV1.

The average number of topic labels per document in RCV1 and RCV2 is about 3, thus we have a multi-label classification problem to solve. We trained and translated binary classifier for each individual topic category, and we evaluated the performance in the F_1 measure (micro-averaged and macro-averaged), which has been conventional in text classification [12].

4.2.2 *Uzbek Dataset*

The internal dataset of Uzbek contains two genres of documents: news articles and discussion forum in both English and Uzbek. Thus we have a binary classification problem. The statistics of each category in English and Uzbek are shown in table 5.

4.2.3 *Cross Validation*

For both RCV1/RCV2 and Uzbek dataset, we used 5-fold cross validation and split both source and target documents into 5 folds. For each run we trained a classification model with 3 folds of training data in the source language, then we evaluated the cross-lingual classification results on one fold of the data (the test set) in the target language. Using the same fold splits, we also trained, validated and tested the classification model directly using the labeled data in the target language, which provides the upper bound performance for model translation. We use MonoTrain to denote this upper bound in section 5.

5. RESULTS ANALYSIS

5.1 Bilingual Dictionary Extension

Figure 2 shows the performance curves of RidgeReg, TransLP, CCA and kNN in dictionary extension for five language pairs, where we used MAP@5 as the metric. For all dictionary sizes and language pairs, TransLP and RidgeReg substantially outperformed CCA and kNN. For larger dictionary size, TransLP further outperformed RidgeReg, while RidgeReg substantially outperformed

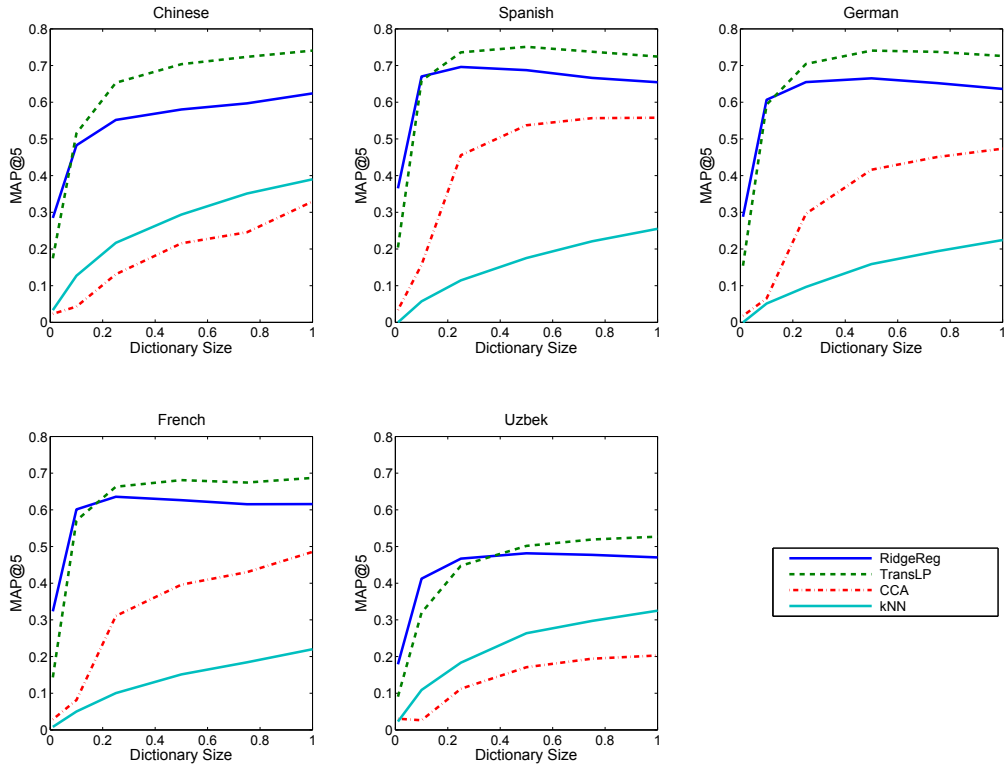

Figure 2: Performance (in MAP@5) of our methods in dictionary extension based on multilingual word embeddings

TransLP for extremely small dictionary size(1% of full size). The performance of RidgeReg got worse as the size of dictionary grows for all languages except for Chinese, this may be due to the fact that we used fixed regularization weight(λ) and slightly suffered from over-fitting for larger dictionary sizes. It is also obvious that the performance of English-Uzbek is worse than four other language pairs. We believe the reason is the relatively smaller size of monolingual corpus(see table 2). In table 3, we present some examples of English words with their predicted translations in other five languages. It could be verified that in most cases the first few extended word pairs are exact translational equivalence. Some predicted translations are very close in meanings, but not identical. For example "style" has been predicted with "baustil"("architecture" in German). It is also interesting to see sometimes the predicted word pairs captured cross-lingual antonyms, like "公用" ("public" in Chinese) was linked with "private". These results are highly informative for understanding the importance of choosing the right method for bridging the language barriers.

5.2 CLTC

Figures 3 and 4 show the end-to-end evaluation results on RCV1/RCV2 of our proposed methods and baselines in simulated low-resource conditions for the four target languages (Chinese, Spanish, German and French) with shared source language (English). Figure 5 shows the result in similar setting for Uzbek dataset. We also include the model translation using non-extended dictionaries, which is named as DictOnly.

Intuitively, the quality of bilingual dictionary should have significant impact on CLTC performance. In general the end-to-end evaluation results are consistent with the ones of dictionary extension. However, there are some inconsistencies between RidgeReg and TransLP and between CCA and kNN on two tasks. For example, for dictionary extension, RidgeReg performed better than TransLP at smaller dictionary size but worse at larger dictionary size. However in Spanish and Uzbek, RidgeReg classified documents better than TransLP at all dictionary sizes. A possible reason would be that for dictionary extension, we require extended word pairs to be exact translational equivalence, while in CLTC it is acceptable to have translations with close but not exactly the same meanings.

On RCV1/RCV2, TransLP and RidgeReg significantly outperformed CCA and kNN, which shows our proposed dictionary extension methods are better suitable for CLTC task than the state-of-art multilingual word embedding techniques and intuitive heuristic. The relative advantage of our methods is even more pronounced when the dictionary sizes are small, which means that these methods are particularly helpful with sparse training data and in low-resource conditions. For the simpler task on Uzbek dataset, RidgeReg outperforms all other methods by a large margin. Again, the advantage is more obvious when the dictionary sizes are small.

Another important observation we could draw from both datasets is that CLNB with RidgeReg, TransLP and kNN all substantially outperformed the results of DictOnly, even when the dictionary is relatively comprehensive. This implies that directly using bilingual dictionaries alone is sub-optimal for CLTC, but using them to establish the mapping among multilingual word embeddings is a winning strategy. For example, with RidgeReg using only 5% to

Table 3: Example English words with their closest words in Chinese(ZH), Spanish(ES), German(DE), French(FR), and Uzbek(UZ), using training results from TransLP at 100% size of bilingual dictionary

Word	Lang	Nearest neighbors	Word	Lang	Nearest neighbors
kill	ZH	死人, 打死, 殺死	research	ZH	實驗, 研討會, 分析, 科研
	ES	matarlo, asesinar, derribar, atrapar		ES	investigaciones, científico, científicos
	DE	ermorden, tötet, zerstören, sterben		DE	untersuchungen, forschungen, untersuchung
	FR	tuer, blesser, tue		FR	bibliographique, recherches, recherche
	UZ	dushman, asir, oldirilgan		UZ	tahlil, ilmiy-tadqiqot, tadqiqot, tadqiqotlar
conference	ZH	會上, 全會, 會, 國務院	private	ZH	公用, 私人, 公有, 私有
	ES	reunión, congresos, agenda, foros		ES	privadas, privados, exclusivo, doméstico
	DE	konferenz, tagung, sitzung		DE	privaten, private, privater, öffentliche
	FR	réunion, forum, colloque, rubrique		FR	privé, privées, privés, exclusif
	UZ	anjuman, yigilish, sessiya, maruza		UZ	ochiq, boshqa, qonuniy, omonatchi
president	ZH	總書記, 澤民, 人知, 孫中山	style	ZH	面貌, 格局, 技法, 手法
	ES	presidida		ES	estilo, gusto, típica, principios
	DE	außenminister, botschafter		DE	baustil, stile, musikstil, spielweise
	FR	president, présidents, l'ambassadeur		FR	mode, style, look, type
	UZ	prezident, qirol, rahbar, prezidenti		UZ	taktika, shakl, tartib, eklektizm

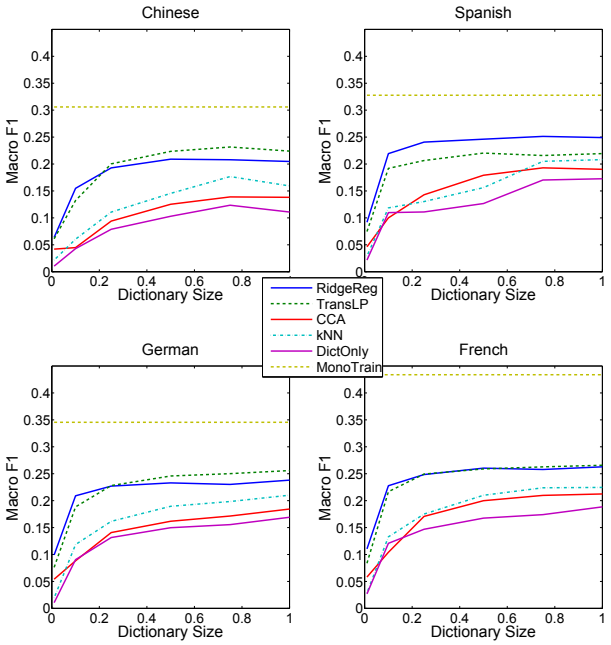

Figure 3: Macro-average F1 curves of our CLTC methods for RCV1/RCV2

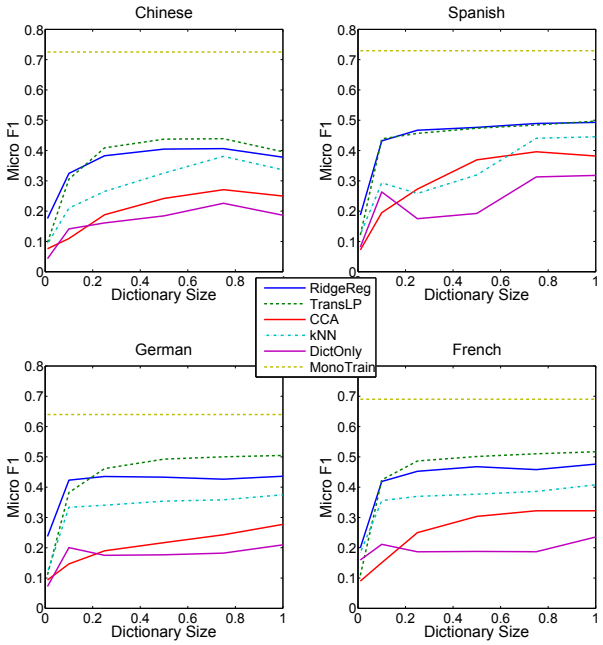

Figure 4: Micro-average F1 curves of our CLTC methods for RCV1/RCV2

10% of the dictionaries, we obtained much higher scores in CLTC than that of DictOnly using 100% of the dictionaries in the experiments for all the language pairs.

Table 6 compares the performance of our methods (CLNB. RidgeReg, CLNB. TransLP) with that of the method (CLMM) by Shi et al. [22] and the method (DR.RidgeReg) originally proposed by Klementiev et al. and followed by Lauly et al. [10, 20]. The experiments were conducted in two datasets and under the condition that each method used the full-sized bilingual dictionaries(i.e. same cross-lingual knowledge). Recall that CLMM was originally evaluated under the conditions of using full-sized human-defined dictionaries, so we conducted this comparative evaluation under the same condition [7]. We also include the performance of kNN and DictOnly for reference. Both our methods performed comparable or better than CLMM in both datasets and in all language pairs, which shows that although designed for low-resource situations, our proposed methods could reach or beat the performance of state-of-art designed for rich-resource scenario. The result is not surprising because our methods are capable of utilizing large and general-purpose monolingual corpus in addition to bilingual dictionary. Another advantage of our methods is efficiency, CLMM is signif-

[7] In our experiments CLMM using extended dictionaries had worse results than CLMM when using non-extended dictionaries.

Table 6: Performance of CLNB, kNN, DictOnly, CLMM, DR using full-sized dictionaries: the results are presented in the format of "Macro-averaged F1/Micro-averaged F1"; bold-face indicates are the best scores for each target language.

Dataset	LANG	CLNB.TransLP	CLNB.RidgeReg	CLMM	DR.RidgeReg	kNN	DictOnly
RCV1/RCV2	Chinese	**0.252**/0.438	0.202/0.399	0.129/0.295	0.059/0.159	0.238/**0.461**	0.098/0.124
	Spanish	0.220/0.506	**0.287**/0.537	0.264/**0.563**	0.099/0.193	0.242/0.475	0.173/0.318
	German	**0.266/0.510**	0.245/0.438	0.220/0.482	0.121/0.295	0.232/0.407	0.169/0.210
	French	**0.267**/0.522	**0.267**/0.487	0.250/**0.553**	0.097/0.286	0.248/0.413	0.188/0.236
Uzbek	Uzbek	0.861/0.886	**0.929/0.950**	0.852/0.877	0.924/0.945	0.885/0.913	0.798/0.803

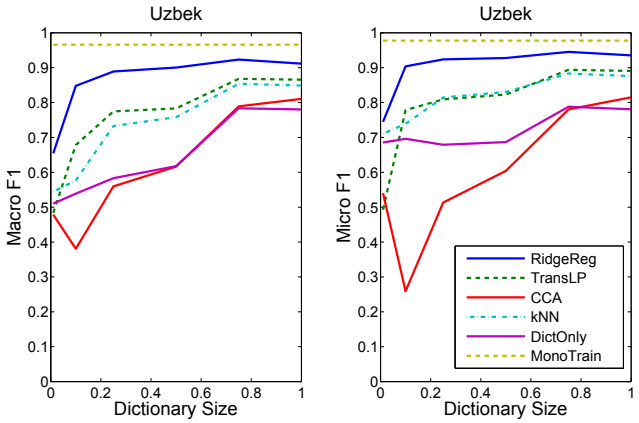

Figure 5: Micro-average and Marco-averaged F1 curves of our CLTC methods for Uzbek dataset

icantly slower than CLNB. TransLP or CLNB. RidgeReg at testing phrase. For CLMM, the prediction time complexity is $\|V\|\times$ (branching factor of dictionary), while the time complexity for CLNB. TransLP and CLNB. RidgeReg is only the averaged number of features(words) for each test data(document). For our setting, CLMM is hundreds of times slower than CLNB, which makes the large-scale evaluation with varying dictionary size intractable for CLMM.

Another interesting comparison is between the different ways of CLTC given multilingual word embeddings. As we could observe from the performance of CLNB.RidgeReg and DR.RidgeReg on RCV1/RCV2, using cross-lingual word similarity to translate the classification model works significantly better than using weighted summation of word vectors as document representation. However, for the simpler task on the Uzbek dataset, the two methods shared similar performance. Therefore, CLNB is always the optimal choice of CLTC, especially for more practical and complicated problems.

6. CONCLUSIONS AND FUTURE WORK

In this paper we present the first study on classification model translation for CLTC under low-resource conditions. We propose a set of new approaches to the extension of bilingual dictionaries via multilingual word embedding and cross-lingual statistical mapping of the embedded words. Experiments on RCV1/RCV2 multilingual document collections and a real low-resource language corpus provide strong evidence for promising performance of our approaches. We hope this study is informative for a new and exciting direction in CLTC research. Future directions include to develop powerful methods for cross-lingual model translation with a broader range of classifiers, and to enable category-sensitive model translation, multi-scale model translation, and multi-scale category-sensitive word embedding.

7. ACKNOWLEDGMENTS

This work is sponsored in part by Defense Advanced Research Projects Agency Information Innovation Office (I2O), the Low Resource Languages for Emergent Incidents (LORELEI) Program, Issued by DARPA/I2O under Contract No. HR0011-15-C-0114, by the National Science Foundation (NSF) under grants IIS-1216282 and IIS-1546329, and by DARPA grant FA8750-12-2-0342 funded under the DEFT program.

8. REFERENCES

[1] M. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in neural information processing systems*, pages 28–36, 2009.

[2] N. Bel, C. H. Koster, and M. Villegas. Cross-lingual text categorization. *Research and Advanced Technology for Digital Libraries*, pages 126–139, 2003.

[3] M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *Conference of the European Chapter of the Association for Computational Linguistics*, page 462–471. Association for Computational Linguistics, 2014.

[4] M. Gardner, K. Huang, E. Papalexakis, X. Fu, P. Talukdar, C. Faloutsos, N. Sidiropoulos, and T. Mitchell. Translation invariant word embeddings. In *Conference on Empirical Methods in Natural Language Processing*, page 1084–1088, 2015.

[5] S. Gouws, Y. Bengio, and G. Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of The 32nd International Conference on Machine Learning*, page 748–756, 2015.

[6] J. Guo, W. Che, D. Yarowsky, H. Wang, and T. Liu. Cross-lingual dependency parsing based on distributed representations. In *Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, page 1234–1244, 2015.

[7] J. Guo, W. Che, D. Yarowsky, H. Wang, and T. Liu. A representation learning framework for multi-source transfer parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[8] Y. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. *arXiv preprint arXiv:1206.6481*, 2012.

[9] Y. Guo and M. Xiao. Transductive representation learning for cross-lingual text classification. In *Data Mining (ICDM),*

2012 IEEE 12th International Conference on, pages 888–893. IEEE, 2012.

[10] A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474, 2012.

[11] T. Kočiskỳ, K. M. Hermann, and P. Blunsom. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*, 2014.

[12] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[13] X. Ling, G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. Can chinese web pages be classified with english data source? In *Proceedings of the 17th international conference on World Wide Web*, pages 969–978. ACM, 2008.

[14] H. Liu and Y. Yang. Bipartite edge prediction via transductive learning over product graphs. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1880–1888, 2015.

[15] A. McCallum, K. Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48, 1998.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[17] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[19] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013.

[20] S. C. A. P, S. Lauly, H. Larochelle, M. M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

[21] L. Rigutini, M. Maggini, and B. Liu. An em based training algorithm for cross-language text categorization. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 529–535. IEEE, 2005.

[22] L. Shi, R. Mihalcea, and M. Tian. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067. Association for Computational Linguistics, 2010.

[23] A. Søgaard, Željko Agić, H. M. Alonso, B. Plank, B. Bohnet, and A. Johannsen. Inverted indexing for cross-lingual nlp. In *Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, page 1713–1722, 2015.

[24] S. Upadhyay, M. Faruqui, C. Dyer, and D. Roth. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, page 1661–1670, 2016.

[25] I. Vulic and M.-F. Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, page 719–725. ACL, 2015.

[26] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics, 2009.

[27] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013.