Cross-lingual Distillation for Text Classification

Ruochen Xu
Carnegie Mellon Universit
ruochenx@cs.cmu.edu

Yiming Yang
Carnegie Mellon Universit
yiming@cs.cmu.edu

Abstract

Cross-lingual text classification(CLTC) is the task of classifying documents written in different languages into the same taxonomy of categories. This paper presents a novel approach to CLTC that builds on model distillation, which adapts and extends a framework originally proposed for model compression. Using soft probabilistic predictions for the documents in a label-rich language as the (induced) supervisory labels in a parallel corpus of documents, we train classifiers successfully for new languages in which labeled training data are not available. An adversarial feature adaptation technique is also applied during the model training to reduce distribution mismatch. We conducted experiments on two benchmark CLTC datasets, treating English as the source language and German, French, Japan and Chinese as the unlabeled target languages. The proposed approach had the advantageous or comparable performance of the other state-of-art methods.

1 Introduction

The availability of massive multilingual data on the Internet makes cross-lingual text classification (CLTC) increasingly important. The task is defined as to classify documents in different languages using the same taxonomy of predefined categories.

CLTC systems build on supervised machine learning require a sufficiently amount of labeled training data for every domain of interest in each language. But in reality, labeled data are not evenly distributed among languages and across domains. English, for example, is a label-rich language in the domains of news stories, Wikipedia pages and reviews of hotels, products, etc. But many other languages do not necessarily have such rich amounts of labeled data. This leads to an open challenge in CLTC, i.e., how can we effectively leverage the trained classifiers in a label-rich *source* language to help the classification of documents in other label-poor *target* languages?

Existing methods in CLTC use either a bilingual dictionary or a parallel corpus to bridge language barriers and to translate classification models (Xu et al., 2016) or text data(Zhou et al., 2016a). There are limitations and challenges in using either type of resources. Dictionary-based methods often ignore the dependency of word meaning and its context, and cannot leverage domain-specific disambiguation when the dictionary on hand is a general-purpose one. Parallel-corpus based methods, although more effective in deploying context (when combined with word embedding in particular), often have an issue of domain mismatch or distribution mismatch if the available source-language training data, the parallel corpus (human-aligned or machine-translation induced one) and the target documents of interest are not in exactly the same domain and genre(Duh et al., 2011). How to solve such domain/distribution mismatch problems is an open question for research.

This paper proposes a new parallel-corpus based approach, focusing on the reduction of domain/distribution matches in CLTC. We call this approach Cross-lingual Distillation with Feature Adaptation or CLDFA in short. It is inspired by the recent work in model compression (Hinton et al., 2015) where a large ensemble model is transformed to a compact (small) model. The assumption of knowledge distillation for model compression is that the knowledge learned by the large model can be viewed as a mapping from in-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1415–1425
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-1130

put space to output (label) space. Then, by training with the soft labels predicted by the large model, the small model can capture most of the knowledge from the large model. Extending this key idea to CLTC, if we see parallel documents as different instantiations of the same semantic concepts in different languages, a target-language classifier should gain the knowledge from a well-trained source classifier by training with the target-language part of the parallel corpus and the soft labels made by the source classifier on the source language side. More specifically, we propose to distillate knowledge from the source language to the target language in the following 2-step process:

- Firstly, we train a source-language classifier with both labeled training documents and adapt it to the unlabeled documents from the source-language side of the parallel corpus. The adaptation enforces our classifier to extract features that are: 1) discriminative for the classification task and 2) invariant with regard to the distribution shift between training and parallel data.

- Secondly, we use the trained source-language classifier to obtain the *soft* labels for a parallel corpus, and the target-language part of the parallel corpus to train a target classifier, which yields a similar category distribution over target-language documents as that over source-language documents. We also use unlabeled testing documents in the target language to adapt the feature extractor in this training step.

Intuitively, the first step addresses the potential domain/distribution mismatch between the labeled data and the unlabeled data in the source language. The second step addresses the potential mismatch between the target-domain training data (in the parallel corpus) and the test data (not in the parallel corpus). The soft-label based training of target classifiers makes our approach unique among parallel-corpus based CLTC methods (Section 2.1. The feature adaptation step makes our framework particularly robust in addressing the distributional difference between in-domain documents and parallel corpus, which is important for the success of CLTC with low-resource languages.

The main contributions in this paper are the following:

- We propose a novel framework (CLDFA) for knowledge distillation in CLTC through a parallel corpus. It has the flexibility to be built on a large family of existing monolingual text classification methods and enables the use of a large amount of unlabeled data from both source and target language.

- CLDFA has the same computational complexity as the plug-in text classification method and hence is very efficient and scalable with the proper choice of plug-in text classifier.

- Our evaluation on benchmark datasets shows that our method had a better or at least comparable performance than that of other state-of-art CLTC methods.

2 Related Work

Related work can be outlined with respect to the representative work in CLTC and the recent progress in deep learning for knowledge distillation.

2.1 CLTC Methods

One branch of CLTC methods is to use lexical level mappings to transfer the knowledge from the source language to the target language. The work by Bel et al. (Bel et al., 2003) was the first effort to solve CLTC problem. They translated the target-language documents to source language using a bilingual dictionary. The classifier trained in the source language was then applied on those translated documents. Similarly, Mihalcea et al. (Mihalcea et al., 2007) built cross-lingual classifier by translating subjectivity words and phrases in the source language into the target language. Shi et al. (Shi et al., 2010) also utilized a bilingual dictionary. Instead of translating the documents, they tried to translate the classification model from source language to target language. Prettenhofer and Stein. (Prettenhofer and Stein, 2010) also used the bilingual dictionary as a word translation oracle and built their CLTC system on structural correspondence learning, a theory for domain adaptation. A more recent work by (Xu et al., 2016) extended seminal bilingual dictionaries with unlabeled corpora in low-resource languages. Chen et al. (Chen et al., 2016) used bilingual word embedding to map documents in source and target language into the same semantic space, and adversarial training was applied to enforce the trained classifier to be language-invariant.

Some recent efforts in CLTC focus on the use of automatic machine translation (MT) technology. For example, Wan (Wan, 2009) used machine translation systems to give each document a source-language and a target-language version, where one version is machine-translated from the another one. A co-training (Blum and Mitchell, 1998) algorithm was applied on two versions of both source and target documents to iterative train classifiers in both languages. MT-based CLTC also include the work on multi-view learning with different algorithms, such as majority voting(Amini et al., 2009), matrix completion(Xiao and Guo, 2013) and multi-view co-regularization(Guo and Xiao, 2012a).

Another branch of CLTC methods focuses on representation learning or the mapping of the induced representations in cross-language settings (Guo and Xiao, 2012b; Zhou et al., 2016a, 2015, 2016b; Xiao and Guo, 2013; Jagarlamudi et al., 2011; De Smet et al., 2011; Vinokourov et al., 2002; Platt et al., 2010; Littman et al., 1998). For example, Meng et al. (Meng et al., 2012) and Lu et al. (Lu et al., 2011) used a parallel corpus to learn word alignment probabilities in a pre-processing step. Some other work attempts to find a language-invariant (or interlingua) representation for words or documents in different languages using various techniques, such as latent semantic indexing (Littman et al., 1998), kernel canonical correlation analysis (Vinokourov et al., 2002), matrix completion(Xiao and Guo, 2013), principal component analysis (Platt et al., 2010) and Bayesian graphical models (De Smet et al., 2011).

2.2 Knowledge Distillation

The idea of distilling knowledge in a neural network was proposed by Hinton et al (Hinton et al., 2015), in which they introduced a student-teacher paradigm. Once the cumbersome teacher network was trained, the student network was trained according to soften predictions of the teacher network. In the field of computer vision, it has been empirically verified that student network trained by distillation performs better than the one trained with hard labels. (Hinton et al., 2015; Romero et al., 2014; Ba and Caruana, 2014). Gupta et al.(Gupta et al., 2015) transfers supervision between images from different modalities(e.g. from RGB image to depth image). There are also some recent works applied distillation in the field of natural language. For example, Lili et al. (Mou et al., 2015) distilled task specific knowledge from a set of high-dimensional embeddings to a low-dimensional space. Zhiting et al. used an iterative distillation method to transfer the structured information of logic rules into the weights of a neural network. Kim et al. (Kim and Rush, 2016) applied knowledge distillation approaches in the field of machine translation to reduce the size of neural machine translation model. Our framework shares the same purpose of existing works that transfer knowledge between models of different properties, such as model complexity, modality, and structured logic. However, our transfer happens between models working on different languages. To the best of knowledge, this is the first work using knowledge distillation to bridge the language gap for NLP tasks.

3 Preliminary

3.1 Task and Notation

CLTC aims to use the training data in the source language to build a model applicable in the target language. In our setting, we have labeled data in source language $L_{src} = \{x_i, y_i\}_{i=1}^{L}$, where x_i is the labeled document in source language and y_i is the label vector. We then have our test data in the target language, given by $T_{tgt} = \{x'_i\}_{i=1}^{T}$. Our framework can also use unlabeled documents from both languages in transductive learning settings. We use $U_{src} = \{x_i\}_{i=1}^{M}$ to denote source-language unlabeled documents,$U_{tgt} = \{x'_i\}_{i=1}^{N}$ to denote target-language unlabeled documents, and $U_{parl} = \{(x_i, x'_i)\}_{i=1}^{P}$ to denote a unlabeled bilingual parallel corpus where x_i and x'_i are paired document translations of each other. We assume that the unlabeled parallel corpus does not overlap with the source-language training documents and the target-language test documents.

3.2 Convolutional Neural Network (CNN) as a Plug-in Classifier

We use a state-of-the-art CNN-based neural network classifier (Kim, 2014) as the plug-in classifier in our framework. Instead of using a bag-of-words representation for each document, the CNN model concatenates the word embeddings (vertical vectors) of each input document into a $n \times k$

matrix, where n is the length (number of word occurrences) of the document, and k is the dimension of word embedding. Denoting by

$$\mathbf{x_{1:n}} = \mathbf{x_1} \oplus \mathbf{x_2} \oplus ... \oplus \mathbf{x_n}$$

as the resulted matrix, with $\oplus$ the concatenation operator. One-dimensional convolutional filter $\mathbf{w} \in R^{hk}$ with window size h operates on every consecutive h words, with non-linear function f and bias b. For window of size h started at index i, the feature after convolutional filter is given by:

$$c_i = f(\mathbf{w} \cdot \mathbf{x_{i:i+h-1}} + b)$$

A max-over-time pooling (Collobert et al., 2011) is applied on c over all possible positions such that each filter extracts one feature. The model uses multiple filters with different window sizes. The concatenated outputs from filters consist the feature of each document. We can see the convolutional filters and pooling layers as feature extractor $\mathbf{f} = G_f(x, \theta_f)$, where θ_f contains parameters for embedding layer and convolutional layer. Theses features are then passed to a fully connected softmax layer to produce probability distributions over labels. We see the final fully connected softmax layer as a label classifier $G_y(\mathbf{f}, \theta_y)$ that takes the output $\mathbf{f}$ from the feature extractor. The final output of model is given by $G_y(G_f(x, \theta_f), \theta_y)$, which is jointly parameterized by $\{\theta_f, \theta_y\}$

We want to emphasize that our choice of the plug-in classifier here is mainly for its simplicity and scalability to demonstrate our framework. There is a large family of neural classifiers for monolingual text classification that could be used in our framework as well, including other convolutional neural networks by (Johnson and Zhang, 2014), the recurrent neural networks by (Lai et al., 2015; Zhang et al., 2016; Johnson and Zhang, 2016; Sutskever et al., 2014; Dai and Le, 2015), the attention mechanism by (Yang et al., 2016), the deep dense network by (Iyyer et al., 2015), and more.

4 Proposed Framework

Let us introduce two versions of our model for cross-language knowledge distillation, i.e., the vanilla version and the full version with feature adaptation. Both are supported by the proposed framework. We denote the former by CLD-KCNN and the latter by CLDFA-KCNN.

4.1 Vanilla Distillation

Without loss of generality, assume we are learning a multi-class classifier for the target language. We have $y \in 1, 2, ..., |v|$ where v is the set of all possible classes. We assume the base classification network produces real number logits q_j for each class. For example, for the case of CNN text classifier, the logits can be produced by a linear transformation which takes features extracted max-pooling layer and outputs a vector of size $|v|$. The logits are converted into probabilities of classes through the softmax layer, by normalizing each q_j with all other logits.

$$p_j = \frac{\exp(q_j/T)}{\sum_{k=1}^{|v|} \exp(q_k/T)} \tag{1}$$

where T is a temperature and is normally set to 1. Using a higher value of T produces a softer probability distribution over classes.

The first step of our framework is to train the source-language classifier on labeled source documents L_{src}. We use standard temperature $T = 1$ and cross-entropy loss as the objective to minimize. For each example and its label (x_i, y_i) from the source training set, we have:

$$\mathcal{L}(\theta_{src}) = \\ -\sum_{(x_i,y_i)\in L_{src}} \sum_{k=1}^{|v|} \mathbb{1}\{y_i = k\} \log p(y = k|x_i; \theta_{src}) \tag{2}$$

where $p(y = k|x; \theta_{src})$ is source model controlled by parameter θ_{src} and $\mathbb{1}\{\cdot\}$ is the indicator function.

In the second step, the knowledge captured in θ_{src} is transferred to the distilled model in the target language by training it on the parallel corpus. The intuition is that paired documents in parallel corpus should have the same distribution of class predicted by the source model and target model. In the simplest version of our framework, for each source-language document in the parallel corpus, we predict a soft class distribution by source model with high temperature. Then we minimize the cross-entropy between soft distribution produced by source model and the soft distribution produced by target model on the paired documents in the target language. More formally, we optimize θ_{tgt} according to the following loss function for each document pair (x_i, x'_i) in parallel corpus.

$$\mathcal{L}(\theta_{tgt}) = - \sum_{(x_i, x'_i) \in U_{parl}} \sum_{k=1}^{|v|} p(y = k | x_i; \theta_{src}) \log p(y = k | x'_i; \theta_{tgt}) \tag{3}$$

During distillation, the same high temperature is used for training target model. After it has been trained, we set the temperature to 1 for testing.

We can show that under some assumptions, the two-step cross-lingual distillation is equivalent to distilling a target-language classifier in the target-language input space.

Lemma 1. *Assume the parallel corpus $\{x_i, x'_i\} \in U_{parl}$ is generated by $x'_i \sim p(X'; \eta)$ and $x_i = t(x'_i)$, where η controls the marginal distribution of x_i and t is a differentiable translation function with integrable derivative. Let $f_{\theta_{src}}(t(x'))$ be the function that outputs soft labels of $p(y = k | t(x'); \theta_{src})$. The distillation given by equation 3 can be interpreted as distillation of a target language classifier $f_{\theta_{src}}(t(x'))$ on target language documents sampled from $p(X'; \eta)$.*

$f_{\theta_{src}}(t(x'))$ is the classifier that takes input of target documents, translates them into source documents through t and makes prediction using the source classifier. If we further assume the testing documents have the same marginal distribution $P(X'; \eta)$, then the distilled classifier should have similar generalization power as $f_{\theta_{src}}(t(x'))$.

Theorem 2. *Let source training data $x_i \in L_{src}$ has marginal distribution $p(X; \lambda)$. Under the assumptions of lemma 1, further assume $p(t(x'); \lambda) = p(x'; \eta)$, $p(y|t(x')) = p(y|x')$ and $t'(x') \approx C$, where C is a constant. Then $f_{\theta_{src}}(t(x'))$ actually minimizes the expected loss in target language data $E_{x' \sim p(X;\eta), y \sim p(Y|x')}[L(y, f(t(x')))]$.*

Proof. By definition of equation 2, $f_{\theta_{src}}(x)$ minimizes the expected loss $E_{x \sim p(X;\lambda), y \sim p(Y|x)}[L(y, f(x))]$, where L is cross-entropy loss in our case. Then we can write

$$\begin{aligned}
& E_{x \sim p(X;\lambda), y \sim p(Y|x)}[L(y, f(x))] \\
= & \int p(x; \lambda) \sum_y p(y|x) L(y, f(x)) dx \\
= & \int p(t(x'); \lambda) \sum_y p(y|t(x')) L(y, f(t(x'))) t'(x') dx' \\
\approx & C \int p(x'; \eta) \sum_y p(y|x') L(y, f(t(x'))) dx' \\
= & C E_{x' \sim p(X;\eta), y \sim p(Y|x')}[L(y, f(t(x')))]
\end{aligned}$$

□

4.2 Distillation with Adversarial Feature Adaptation

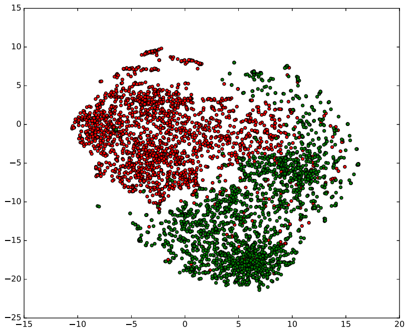


Figure 1: Extracted features for source-language documents in the English-Chinese Yelp Hotel Review dataset. Red dots represent features of the documents in L_{src} and green dots represent the features of documents in U_{parl}, which is a general-purpose parallel corpus.

Although vanilla distillation is intuitive and simple, it cannot handle distribution mismatch issues. For example, the marginal feature distributions of source-language documents in L_{src} and U_{parl} could be different, so are the distributions of target-language documents in U_{parl} and T_{tgt}. According to theorem 2, the vanilla distillation works for the best performance under unrealistic assumption: $p(t(x')|\lambda) = p(x'|\eta)$. To further illustrate our point, we trained a CNN classifier according to equation 2 and used the features extracted by G_f to present the source-language documents in both L_{src} and U_{parl}. Then we projected the high-dimensional features onto a 2-dimensional space via t-Distributed Stochastic Neighbor Embedding (t-SNE)(Maaten and Hinton, 2008). This resulted the visualization of the project data in Figures 1 and 2.

It is quite obvious in Figure 1 that the general-purpose parallel corpus has a very different feature distribution from that of the labeled source training set. Even for machine-translated parallel data from the same domain, as shown in figure 2, there is still a non-negligible distribution shift from the source language to the target language for the extracted features. Our interpretation of this observation is that when the MT system (e.g. Google Translate) is a general-purpose one, it non-avoidably add translation ambiguities which would lead the distribution shift from the original domain. To address the distribution divergence brought by either a general-purpose parallel corpus or an imperfect MT system, we seek to adapt the features extraction part of our neural classifier such that the feature distributions on both sides should be close as possible in the newly induced feature space. We adapt the adversarial training method by (Ganin and Lempitsky, 2014) to the cross-lingual settings in our problems.

Given a set of training set of $L = \{x_i, y_i\}_{i=1,\dots,N}$ and an unlabeled set $U = \{x'_i\}_{i=1,\dots,M}$, our goal is to find a neural classifier $G_y(G_f(x, \theta_f), \theta_y)$, which has good discriminative performance on L and also extracts features which have similar distributions on L and U. One way to maximize the similarity of two distributions is to maximize the loss of a discriminative classifier whose job is to discriminate the two feature distributions. We denote this classifier by $G_d(\cdot, \theta_d)$, which is parameterized by θ_d.

At training time, we seek θ_f to minimize the loss of G_y and maximize the loss of G_d. Meanwhile, θ_y and θ_d are also optimized to minimize their corresponding loss. The overall optimization could be summarized as follows:

$$E(\theta_f, \theta_y, \theta_d) = \sum_{x_i, y_i \in L} L_y(y_i, G_y(G_f(x_i, \theta_f), \theta_y))$$
$$- \alpha \sum_{x_i \in L} L_d(0, G_d(G_f(x_i, \theta_f), \theta_d))$$
$$- \alpha \sum_{x_j \in U} L_d(1, G_d(G_f(x_j, \theta_f), \theta_d))$$

where L_y is the loss function for true labels y, L_d is loss function for binary labels indicating the source of data and α is the hyperparameter that controls the relative importance of two losses. We optimize θ_f, θ_y for minimizing E and optimize θ_d for maximizing E. We jointly optimize $\theta_f, \theta_y, \theta_d$ through the gradient reversal layer(Ganin and Lempitsky, 2014).

We use this feature adaptation technique to firstly adapt the source-language classifier to the source-language documents of the parallel corpus. When training the target-language classifier by matching soft labels on the parallel corpus, we also adapt the classifier to the target testing documents. We use cross-entropy loss functions as L_y and L_d for both feature adaptation.

5 Experiments and Discussions

5.1 Dataset

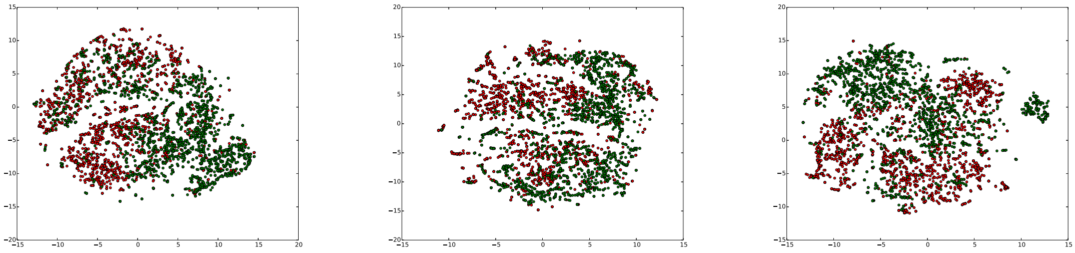
Our experiments used two benchmark datasets, as described below.

(1) Amazon Reviews

Language	Domain	# of Documents
English	book	50000
	DVD	30000
	music	25220
German	book	165470
	DVD	91516
	music	60392
French	book	32870
	DVD	9358
	music	15940
Japanese	book	169780
	DVD	68326
	music	55892

Table 1: Dataset Statistics for the Amazon reviews dataset

We used the multilingual multi-domain Amazon review dataset created by Prettenhofer and Stein (Prettenhofer and Stein, 2010). The dataset contains Amazon reviews in three domains: book, DVD and music. Each domain has the reviews in four different languages: English, German, French and Japanese. We treated English as the source language and the rest three as the target languages, respectively. This gives us 9 tasks (the product of the 3 domains and the 3 target languages) in total. For each task, there are 1000 positive and 1000 negative reviews in English and the target language, respectively. (Prettenhofer and Stein, 2010) also provides 2000 parallel reviews per task,



(a) Germany:DVD

(b) French:Music

(c) Japanese:Book

Figure 2: Extracted features for the source-language documents in the Amazon Reviews dataset. Red dots represent the features of the labeled training documents in L_{src}, and green dots represent the features of the documents in U_{parl}, which are the machine-translated documents from a target language. Below each figure is the target language and the domain of review (Section 5.1).

that were generated using Google Translate [1], and used by us for cross-language distillation. There are also several thousands of unlabeled reviews in each language. The statistics of unlabeled data is summarized in Table 1. All the reviews are tokenized using standard regular expressions except for Japanese, for which we used a publicly available segmenter [2].

(2) English-Chinese Yelp Hotel Reviews

This dataset was firstly used for CLTC by (Chen et al., 2016). The task is to make sentence-level sentiment classification with 5 labels(rating scale from 1 to 5), using English as the source language and Chinese as the target language. The labeled English data consists of balanced labels of 650k Yelp reviews from Zhang et al. (Zhang et al., 2015). The Chinese data includes 20k labeled Chinese hotel reviews and 1037k unlabeled ones from (Lin et al., 2015). Following the approach by (Chen et al., 2016), we use 10k of labeled Chinese data as validation set and another 10k hotel reviews as held-out test data. We a random sample of 500k parallel sentences from UM-courpus(Tian et al., 2014), which is a general-purpose corpus designed for machine translation.

5.2 Baselines

We compare the proposed method with other state-of-the-art methods as outlined below.

(1) Parallel-Corpus based CLTC Methods

Methods in this category all use an unlabeled parallel corpus. Methods named **PL-LSI** (Littman et al., 1998), **PL-OPCA** (Platt et al., 2010) and **PL-KCAA** (Vinokourov et al., 2002) learn latent document representations in a shared low-dimensional space by performing the Latent Semantic Indexing (LSI), the Oriented Principal Component Analysis (OPCA) and a kernel (namely KCAA) for the parallel text. **PL-MC** (Xiao and Guo, 2013) recovers missing features via matrix Completion, and also uses LSI to induce a latent space for parallel text. All these methods train a classifier in the shared feature space with labeled training data from both the source and target languages.

(2) MT-based CLTC Methods

The methods in this category all use an MT system to translate each test document in the target language to the source language in the testing phase. The prediction on each translated document is made by a source-language classifier, which can be a Logistic Regression model (**MT+LR**) (Chen et al., 2016) or a deep averaging network (**MT+DAN**) (Chen et al., 2016).

(3) Adversarial Deep Averaging Network

Similar to our approach, the adversarial Deep Averaging Network (**ADAN**) also exploits adversarial training for CLTC (Chen et al., 2016). However, it does not have the parallel-corpus based knowledge distillation part (which we do). Instead, it uses averaged bilingual embeddings of words as its input and adapts the feature extractor to produce similar features in both languages.

We also include the results of **mSDA** for the Yelp Hotel Reviews dataset. **mSDA** (Chen et al., 2012) is a domain adaptation method based on

[1] translate.google.com

[2] https://pypi.python.org/pypi/tinysegmenter

Target Language	Domain	PL-LSI	PL-KCCA	PL-OPCA	PL-MC	CLD-KCNN	CLDFA-KCNN
German	book	77.59	79.14	74.72	79.22	82.54	**83.95***
	DVD	79.22	76.73	74.59	81.34	82.24	**83.14***
	music	73.81	79.18	74.45	**79.39**	74.65	79.02
French	book	79.56	77.56	76.55	81.92	81.6	**83.37**
	DVD	77.82	78.19	70.54	81.97	82.41	**82.56**
	music	75.39	78.24	73.69	79.3	83.01	**83.31***
Janpanese	book	72.68	69.46	71.41	72.57	74.12	**77.36***
	DVD	72.55	74.79	71.84	76.6	79.67	**80.52***
	music	73.44	73.54	74.96	76.21	73.69	**76.46**
Averaged Accuracy		75.78	76.31	73.64	78.72	79.33	**81.08***

Table 2: Accuracy scores of methods on the Amazon Reviews dataset: the best score in each row (a task) is highlighted in bold face. If the score of CLDFA-KCNN is statistically significantly better (in one-sample proportion tests) than the best among the baseline methods, it is marked using a star.

Model	Accuracy
mSDA	31.44%
MT-LR	34.01%
MT-DAN	39.66%
ADAN	41.04%
CLD-KCNN	40.96%
CLDFA-KCNN	**41.82%**

Table 3: Accuracy scores of methods on the English-Chinese Yelp Hotel Reviews dataset

stacked denoising autoencoders, which has been proved to be effective in cross-domain sentiment classification evaluations. We show the results reported by (Chen et al., 2012), where they used bilingual word embedding as input for mSDA.

5.3 Implementation Detail

We pre-trained both the source and target classifier with unlabeled data in each language. We ran word2vec(Mikolov et al., 2013) [3] on the tokenized unlabeled corpus. The learned word embeddings are used to initialize the word embedding look-up matrix, which maps input words to word embeddings and concatenates them into input matrix.

We fine-tuned the source-language classifier on the English training data with 5-fold cross-validation. For English-Chinese Yelp-hotel review dataset, the temperature T(Section 4.1) in distillation is tuned on validation set in the target language. For Amazon review dataset, since there is no default validation set, we set temperature from low to high in $\{1, 3, 5, 10\}$ and take the average among all predictions.

5.4 Main Results

In tables 2 and 3 we compare the results of our methods (the vanilla version CLD-KCNN and the full version CLDFA-KCNN) with those of other methods based on the published results in the literature. The baseline methods are different in these two tables as they were previously evaluated (by their authors) on different benchmark datasets. Clearly, CLDFA-KCNN outperformed the other methods on all except one task in these two datasets, showing that knowledge distillation is successfully carried out in our approach. Noticing that CLDFA-KCNN outperformed CLD-KCNN, showing the effectiveness of adversarial feature extraction in reducing the distribution mismatch between the parallel corpus and the train/test data in the target domain. We should also point out that in Table 2, the four baseline methods (PL-LSI, PL-KCCA, PL-OPCA and PL-MC) were evaluated under the condition of using additional 100 labeled target documents for training, according to the author's report (Xiao and Guo, 2013). On the other hand, our methods (CLD-KCNN and CLDFA-KCNN) were evaluated under a tougher condition, i.e., not using any labeled data in the target domains.

We also test our framework when a few training documents in the target language are available. A simple way to utilize the target-language supervision is to fit the target-language model with labeled target data after optimizing with our cross-lingual distillation framework. The performance of CLD-KCNN and CLDFA-KCNN trained with different sizes of labeled target-language data is shown in figure 3. We also compare the performance of training the same classifier using only

[3] https://code.google.com/archive/p/word2vec/

the target-language labels(**Target Only** in figure 3). As we can see, our framework can efficiently utilize the extra supervision and improve the performance over the training using only the target-language labels. The margin is most significant when the size of the target-language label is relatively small.

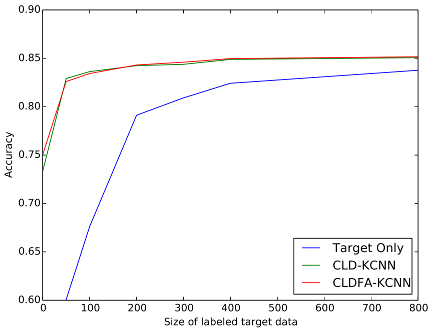


Figure 3: Accuracy scores of methods using varying sizes of target-language labeled data on the Amazon review dataset. The target language is German and the domain is music. The parallel corpus has a fixed size of 1000 and the size of the labeled target-language documents is shown on the x-axis

6 Conclusion

This work introduces a novel framework for distillation of discriminative knowledge across languages, providing effective and efficient algorithmic solutions for addressing domain/distribution mismatch issues in CLTC. The excellent performance of our approach is evident in our evaluation on two CLTC benchmark datasets, compared to that of other state-of-the-art methods.

Acknowledgement

We thank the reviewers for their helpful comments. This work is supported 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 grant IIS-1546329.

References

Massih Amini, Nicolas Usunier, and Cyril Goutte. 2009. Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in neural information processing systems*. pages 28–36.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in neural information processing systems*. pages 2654–2662.

Nuria Bel, Cornelis HA Koster, and Marta Villegas. 2003. Cross-lingual text categorization. *Research and Advanced Technology for Digital Libraries* pages 126–139.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, pages 92–100.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683* .

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Weinberger, and Claire Cardie. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614* .

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*. pages 3079–3087.

Wim De Smet, Jie Tang, and Marie-Francine Moens. 2011. Knowledge transfer across multilingual corpora via latent topics. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pages 549–560.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 429–433.

Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* .

Yuhong Guo and Min Xiao. 2012a. Cross language text classification via subspace co-regularized multiview learning. *arXiv preprint arXiv:1206.6481* .

Yuhong Guo and Min Xiao. 2012b. Transductive representation learning for cross-lingual text classification. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, pages 888–893.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2015. Cross modal distillation for supervision transfer. *arXiv preprint arXiv:1507.00448* .

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the Association for Computational Linguistics*.

Jagadeesh Jagarlamudi, Raghavendra Udupa, Hal Daumé III, and Abhijit Bhole. 2011. Improving bilingual projections via sparse covariance matrices. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 930–940.

Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058* .

Rie Johnson and Tong Zhang. 2016. Supervised and semi-supervised text categorization using lstm for region embeddings. In *Proceedings of The 33rd International Conference on Machine Learning*. pages 526–534.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947* .

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*. pages 2267–2273.

Yiou Lin, Hang Lei, Jia Wu, and Xiaoyu Li. 2015. An empirical study on sentiment classification of chinese review using word embedding. *arXiv preprint arXiv:1511.01665* .

Michael L Littman, Susan T Dumais, and Thomas K Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-language information retrieval*, Springer, pages 51–62.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 320–330.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 572–581.

Rada Mihalcea, Carmen Banea, and Janyce M Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Lili Mou, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2015. Distilling word embeddings: An encoding approach. *arXiv preprint arXiv:1506.04488* .

John C Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual document representations from discriminative projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 251–261.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1118–1127.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* .

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Liang Tian, Derek F Wong, Lidia S Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *LREC*. pages 1837–1842.

Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. 2002. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS*. volume 1, page 4.

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

Min Xiao and Yuhong Guo. 2013. A novel two-step method for cross language representation learning. In *Advances in Neural Information Processing Systems*. pages 1259–1267.

Ruochen Xu, Yiming Yang, Hanxiao Liu, and Andrew Hsi. 2016. Cross-lingual text classification via model translation with limited dictionaries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, pages 95–104.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* .

Rui Zhang, Honglak Lee, and Dragomir Radev. 2016. Dependency sensitive convolutional neural networks for modeling sentences and documents. *arXiv preprint arXiv:1611.02361* .

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*. pages 649–657.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. ACL.

Xinjie Zhou, Xianjun Wan, and Jianguo Xiao. 2016a. Cross-lingual sentiment classification with bilingual document representation learning .

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Attention-based lstm network for cross-lingual sentiment classification .