Cross-lingual Dataless Classification for Many Languages

Yangqiu Song[1] **and Shyam Upadhyay**[2] **and Haoruo Peng**[2] **and Dan Roth**[2]
[1]Lane Department of CSEE, West Virginia University
[2]Department of Computer Science, University of Illinois at Urbana-Champaign
[1]yangqiu.song@mail.wvu.edu, [2]{upadhya3,hpeng7,danr}@illinois.edu

Abstract

Dataless text classification [Chang *et al.*, 2008] is a classification paradigm which maps documents into a given label space without requiring any annotated training data. This paper explores a cross-lingual variant of this paradigm, where documents in multiple languages are classified into an English label space. We use CLESA (cross-lingual explicit semantic analysis) to embed both foreign language documents and an English label space into a shared semantic space, and select the best label(s) for a document using the similarity between the corresponding semantic representations. We illustrate our approach by experimenting with classifying documents in 88 different languages into the same English label space. In particular, we show that CLESA is better than using a monolingual ESA on the target foreign language and translating the English labels into that language. Moreover, the evaluation on two benchmarks, TED and RCV2, showed that cross-lingual dataless classification outperforms supervised learning methods when a large collection of annotated documents is not available.

1 Introduction

Text classification is a fundamental problem in many natural language processing and data mining applications including topic detection and tracking and event extraction and identification. Traditional approaches use supervised learning methods to train a classifier to predict text categories. However, supervised learning, which is difficult to scale even in English, is unrealistic when applied to many other languages, where annotated corpora do not exist. Therefore, besides traditional semi-supervised learning [Chapelle *et al.*, 2006] and transfer learning [Pan and Yang, 2010], cross-lingual document classification was recently proposed as a way to use training data in one language to classify the documents in another language [Amini and Goutte, 2010; Klementiev *et al.*, 2012]. Existing cross-lingual document classification approaches either need a parallel corpus to train word embedding for different languages [Hermann and Blunsom, 2014], require labeled documents in both source and target languages [Xiao and Guo, 2013], make use of machine translation techniques to translate words [Prettenhofer and Stein, 2010] or documents [Amini and Goutte, 2010], or combine different approaches [Shi *et al.*, 2010]. Thus they first rely on the existence of a large parallel corpus (or certain type of alignment) to learn translation or embedding models. They also train the supervised learning algorithms on the labeled documents in source language, and thus still rely heavily on human annotations. However, in practice, both a parallel corpus and labeled documents in a target language are expensive to obtain. Another difficulty for existing methods is that they are not flexible in choosing the categories. By changing from one category space to another, they will need more labeled documents or other forms of human effort.

The above approaches ignore the fact that the labels or the short descriptions of the categories as texts themselves are meaningful. In this paper, we examine the scenario where we are presented with documents written in a foreign language which we don't understand. Nevertheless, we would like to know the topics of these documents and map them to a category space with English short descriptions. Often in such situations, we do not have a good translation model at our disposal. We want the cheapest way to understand the document topics with the flexibility of choosing the label space. The labels in English will then help us classify the documents in other languages, which bypasses the requirements of document translation and labeling.

To address the problem of directly classifying documents in other languages into English label space, we present cross-lingual dataless classification based on cross-lingual explicit semantic analysis (CLESA) [Potthast *et al.*, 2008; Sorg and Cimiano, 2012], to generate a common semantic representation for English labels and foreign language documents. CLESA is a generalization of explicit semantic analysis (ESA) for English [Gabrilovich and Markovitch, 2009], introduced in the context of Information Retrieval [Potthast *et al.*, 2008; Sorg and Cimiano, 2012] and used also for Twitter message classification [Shirakawa *et al.*, 2014]. We extend the ESA approach for monolingual dataless classification [Chang *et al.*, 2008; Song and Roth, 2014] to support multi-lingual representations. In particular, CLESA aligns Wikipedia pages with the same title across languages using language links. By working in this aligned space, CLESA embeds texts in two languages into the same semantic space.

We develop a cross-lingual classification method which requires no parallel corpus or labeled documents, and can classify documents in many languages on the fly. This reduces the task of classifying foreign documents to simply comparing similarities between representations of documents and labels in this common semantic space.

To show the wide applicability of our approach, we evaluate on three multi-lingual classification corpora. We first generate a multi-lingual classification data set across 88 languages by selecting 100 documents from the 20-newsgroups data set [Lang, 1995] and translating them using Google translation into 88 languages. We also use two standard benchmark data sets, TED [Hermann and Blunsom, 2014] and RCV2 (a multi-lingual version of RCV1 [Lewis *et al.*, 2004] for English), to evaluate the cross-lingual dataless classification. We demonstrate the superiority of CLESA by comparing against the approach of performing monolingual ESA with translated labels. Our experiments also show that cross-lingual dataless classification is preferable to the supervised learning methods in the absence of a large collection of annotated documents.

2 Cross-lingual Dataless Classification

In this section, we present the general dataless classification framework and then explain how we extend it to support cross-lingual dataless classification.

2.1 Dataless Classification

Dataless classification performs a nearest neighbor search of labels for a document in an appropriately selected semantic space [Chang *et al.*, 2008; Song and Roth, 2014]. Let $\phi(d)$ be the representation of document d in a semantic space (to be defined later) and let $\{\boldsymbol{\phi}(l^{(1)}), \ldots, \boldsymbol{\phi}(l^{(N_l)})\}$ be the representations of the N_l labels in the same space. Then we can evaluate similarity using an appropriate metric $f(\boldsymbol{\phi}(d), \boldsymbol{\phi}(l^{(i)}))$, (e.g., cosine similarity between two sparse vectors) and select label(s) that maximizes the similarity: $l^* = \arg\max_i f(\boldsymbol{\phi}(d), \boldsymbol{\phi}(l^{(i)}))$.

The core problem in dataless classification is to find a semantic space that enables good representations of documents and labels. Traditional text classification makes use of a bag-of-words (BOW) representation of documents. However, when comparing labels and documents in dataless classification, the brevity of labels makes this simple minded representation and the resulting similarity measure unreliable. For example, a document talking about "sports" does not necessarily contain the word "sports." Consequently, other more expressive distributional representations have been applied, e.g., Brown cluster [Brown *et al.*, 1992; Liang, 2005], neural network embedding [Collobert *et al.*, 2011; Turian *et al.*, 2010; Mikolov *et al.*, 2013b; 2013a], topic modeling [Blei *et al.*, 2003], ESA [Gabrilovich and Markovitch, 2009], and their combinations [Song and Roth, 2015]. It has been shown that ESA gives the best and most robust results for dataless classification for English documents [Song and Roth, 2014]. Thus, we make use of ESA to develop expressive semantic representations across multiple languages.

ESA uses Wikipedia as external world knowledge to generate a set of *titles* for a given fragment of text [Gabrilovich and Markovitch, 2009]. Each word in a text is represented as a weighted vector of the Wikipedia titles in which it is mentioned. This can be computed using an inverted index for each word in Wikipedia. The text fragment representation is then the sum of the IDF (inverse document frequency) weighted vectors that correspond to the words in the text fragment.

2.2 CLESA

In order to support cross-lingual dataless classification, we implemented a version of CLESA [Potthast *et al.*, 2008; Sorg and Cimiano, 2012] that is used for dataless classification scheme by exploiting the shared semantic space between two languages. To build connections between languages, we extract language links from Wikipedia dumps. Each language link shows that a title in one language can also be described in another language. Even though these may not be direct translations, they define the same semantic concept. For example, a Wikipedia page titled "Basketball" has a corresponding Italian page "Pallacanestro," a Spanish page "Baloncesto," etc. Thus, we can intersect the Wikipedia title space of any two languages and use the set of shared Wikipedia titles as the semantic space for texts in both languages. Note that not all intersections of the Wikipedia title spaces give us satisfactory and meaningful semantic spaces. Generally speaking, we find that the larger the intersection, the better the quality of the semantic space is.

Formally, assume that we have Wikipedia dumps for languages A and B. Traditional ESA uses the sparse vector $\boldsymbol{\phi}^A(w_A) = (\phi^A_{C_{1_A}}(w_A), \ldots, \phi^A_{C_{N_A}}(w_A))^T \in \mathbb{R}^{N_A}$ to represent a word w_A where N_A is the number of titles in the language A Wikipedia, and $\phi^A_{C_{i_A}}(w_A)$ is the weight indicating how important word w_A is in the Wikipedia page titled C_i in language A. Similarly, $\boldsymbol{\phi}^B(w_B) = (\phi^B_{C_{1_A}}(w_B), \ldots, \phi^B_{C_{N_B}}(w_B))^T \in \mathbb{R}^{N_B}$ for language B. To compare text similarities between languages A and B, a natural way is to consider first the intersection of the two title sets:

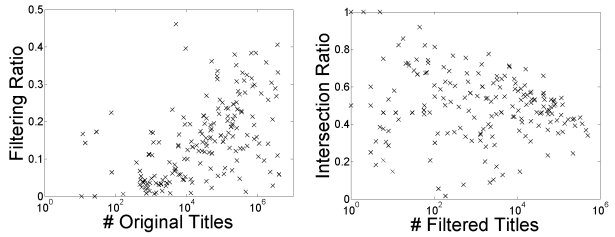
$$\{C_1, \ldots, C_N\} = \{C_{1_A}, \ldots, C_{N_A}\} \bigcap \{C_{1_B}, \ldots, C_{N_B}\}. \quad (1)$$

Thus, we unify the vector representations in A and B as follows:

$$\begin{aligned} \boldsymbol{\phi}(w_A) &\doteq (\phi^A_{C_1}(w_A), \ldots, \phi^A_{C_N}(w_A))^T \in \mathbb{R}^N, \\ \boldsymbol{\phi}(w_B) &\doteq (\phi^B_{C_1}(w_B), \ldots, \phi^B_{C_N}(w_B))^T \in \mathbb{R}^N. \end{aligned} \quad (2)$$

Now, given a document d_A in language A as a vector $(w_{A_1}, \ldots, w_{A_{M_A}})^T \in \mathbb{R}^{M_A}$, where M_A is the vocabulary size of language A. Denote p_{A_i} as the weight of word w_{A_i} in the document. For example, the weight could be TF-IDF, where TF represents the term frequency of word w_{A_i} in d_A, and IDF, the inverse document frequency in Wikipedia. Then we can define the vector representation for d_A as:

$$\boldsymbol{\phi}(d_A) = \frac{1}{M_A} \sum_i p_{A_i} \boldsymbol{\phi}(w_{A_i}). \quad (3)$$



(a) Filtering ratio of with 100 words and 5 links.

(b) Intersection ratio of Wikipedia language links.

Figure 1: The effects of preprocessing of Wikipedia. Each cross in the figures represents a language. (For English Wikipedia, originally we had around 15-million titles. After filtering, we had about 3 million titles.)

Similarly, for a label $l_B^{(i)}$ in language B, we have the vector representation:

$$\boldsymbol{\phi}(l_B^{(i)}) = \frac{1}{M_B} \sum_j p_{B_j}^{(i)} \boldsymbol{\phi}(w_{B_j}^{(i)}), \tag{4}$$

where $l_B^{(i)} = (w_{B_1}^{(i)}, \ldots, w_{B_{M_B}}^{(i)})^T \in \mathbb{R}^{M_B}$ is a highly sparse vector, M_B is the vocabulary size of language B and $p_{B_j}^{(i)}$ represents the weight of word j in the label description $l_B^{(i)}$. Now we can use the cosine similarity between $\boldsymbol{\phi}(d_A)$ and $\boldsymbol{\phi}(l_B^{(i)})$ as in traditional ESA in order to choose the best label: $l^* = \arg\max_i \cos(\boldsymbol{\phi}(d_A), \boldsymbol{\phi}(l_B^{(i)}))$.

2.3 Bootstrapping

Similarly to the bootstrapping approach used in monolingual classification, we use the cross-lingual dataless classification procedure described above as an initialization, and then incrementally label more pseudo-labels to train a supervised classifier to label more data. Our bootstrapping procedure follows [Song and Roth, 2014]:

Step 1: Initialize n documents for each label by using confident CLESA classifications.

Step 2: For each iteration, train a classifier based on BOW representation[1] to label n' more documents for each label. For imbalanced data, we can also set a threshold δ to stop adding more pseudo-labels.

Step 3: Continue until we label all documents.

3 Using CLESA for Dataless Classification

In this section, we first show how we build CLESA representations in many languages. We also present and compare an alternative way to do cross-lingual dataless classification: translating labels into each target language and then applying monolingual dataless classification in the target language. We note that this alternative shares some advantages with our proposed CLESA based method: it does not require heavy resources in the target languages (only the label space is to be translated) as do the other methods we mentioned earlier. However, when we compare this naive method with our proposed approach on a multi-lingual classification data set, it turns out that our CLESA representation is the better choice for many language pairs while also being cheaper in terms of acquiring resources (no translation is needed). Note that we do not compare with another naive approach which translates both documents and labels to English and performs English ESA. This is because: (1) In practice, translation of documents is more costly than translation of labels and requires significantly more resources (label translation can be done once by an expert); (2) There is no large collection of documents in different languages labeled in the same label space to facilitate a fair comparison.

Table 1: Statistics of 179 Wikipedia corpora. "Filtered" corresponds to the numbers of pages after filtering with 100 words and 5 links. This is the data used for monolingual ESA. "$L \cap$ English" corresponds to the CLESA data.

# Titles	# Languages		
	Original	Filtered	$L \cap$ English
$n \geq 10^6$	17	2	0
$10^6 > n \geq 10^5$	50	27	14
$10^5 > n \geq 10^4$	45	44	48
$10^4 > n \geq 10^3$	39	41	41
$10^3 > n \geq 10^2$	21	25	31
$10^2 > n \geq 10$	7	31	24
$10 > n \geq 0$	0	9	21

3.1 Building CLESA Representations

We first downloaded the complete Wikipedia corpus that is available in 180 languages including English.[2] The pages were tokenized and cleaned using the 38 available Lucene[3] language-dependent tokenizers[4] and with a whitespace based tokenizer for other languages. We filtered out pages with fewer than 100 words or 5 language links. This way, most of the redirection and disambiguation pages were removed and some of the short pages were also removed.

We fixed the English label space. Suppose that the target documents are in a foreign language L; in order to map the documents and labels to a common semantic space, we compute the intersection of the Wikipedia title pages linked between English and L. That is, for each language L we only keep those Wikipedia pages that are linked to the English Wikipedia. This results in further reducing the size of the collection available in each language.

Table 1 shows statistics about numbers of titles in the original 179 languages excluding English, after filtering and after intersection with English. There are 62 languages with more than 10,000 Wikipedia titles that are linked to the English Wikipedia. For these 62 language we therefore have a title space that covers a wide range of topics. In Figure 1(a) we show the ratio of remaining titles in each language after filtering short and non-linked pages (threshold=5). This indicates that larger Wikipedias also tend to have longer and higher quality content. In Figure 1(b), the ratio after intersecting

[1] In practice, we found this representation worked best for the data used in this paper.

[2] https://dumps.wikimedia.org/

[3] https://lucene.apache.org/

[4] Stop words are embedded.

Table 2: Precision statistics of 20-newsgroups classification in 88 languages. The numbers in the four columns represent the number of languages among the 88 for which the precision values fall within the ranges indicated on the left. MONO. stands for monolingual ESA. CROSS. stands for cross-lingual ESA.

	Top-1 Precision		Top-3 Precision	
	MONO.	CROSS.	MONO.	CROSS.
$1 \geq p \geq 0.9$	2	20	5	28
$0.9 > p \geq 0.8$	7	8	9	8
$0.8 > p \geq 0.7$	10	7	16	8
$0.7 > p \geq 0.6$	14	4	13	8
$0.6 > p \geq 0.5$	8	9	8	5
$0.5 > p \geq 0.4$	6	5	4	7
$0.4 > p \geq 0.3$	10	9	10	6
$0.3 > p \geq 0.2$	6	9	6	5
$0.2 > p \geq 0.1$	7	8	5	4
$0.1 > p \geq 0$	18	9	12	9

with the English Wikipedia shows that larger-size Wikipedias have a more stable fraction of titles that are linked to the English Wikipedia, relative to smaller-size Wikipedias.

3.2 Monolingual ESA vs. CLESA

For cross-lingual document classification, a natural idea is to translate the documents and perform monolingual ESA. However, translation can be very costly and is not scalable to a large amount of documents. Another option is to keep the original set of Wikipedia titles in language L, and map the English label space to language L. This can be achieved with a relatively small effort compared with translating the documents, since the label space is rather small (i.e., no more than a few hundreds of words). Once we do that, we can generate an ESA representation in L, and run a monolingual dataless classification in L.

In order to understand the difference and relative advantages of this method and the one we proposed and presented earlier in Section 3.1, we perform the following experiment. We first translate a set of English documents to many languages, via Google Translation. (Note that we do this only to generate a new data set on which we can perform a fair comparisons of the algorithms). We then perform the experiment as described above: translating the labels, and developing a monolingual ESA representation in language L, which is then used for dataless classification in language L. Specifically, we select 100 documents from the 20-newsgroups data set [Lang, 1995] which can be correctly classified using the English ESA. Then we use Google Translation API[5] to translate these documents into 88 languages.[6]

Now we can compare two settings for performing dataless classification: using monolingual ESA and using CLESA. We show the results of "top-1 label hit" and "top-3 labels hit" precisions in Table 2. Top-1 label means that for each document, we select the best label to classify it. This is exact classification evaluation. While for Top-3 labels, we select the best three labels for each document and check whether they contain the correct label. Comparing monolingual ESA and CLESA, the results in Table 2 show that even though the number of titles (Wikipedia pages) used by CLESA is much smaller than monolingual ESA, CLESA produces, on average, more accurate classifications. The language links used by CLESA help to disambiguate some Wikipedia titles. For example, some entities such as "python" have multiple meanings, which are better disambiguated when considering multiple languages. Therefore, the shared semantic space generated by CLESA provides a better representation than the single language title space.

4 Benchmark Evaluation

Section 3 established that the use of a common semantic space is a better way to perform dataless classification, and therefore we evaluate our proposed CLESA method in the standard document classification task. We present benchmark results for cross-lingual dataless classification on two data sets, TED and RCV2. Before discussing the results, we first list the baselines and our comparison methodology.

4.1 Experimental Comparison

Our goal is to evaluate the quality of classifying documents in Language L into an English label space. To understand the advantages and shortcomings of our method, we compare the following approaches.

Cross-lingual ESA and bootstrapping. We use the cross-lingual ESA described in Section 2.2 for dataless classification for both data sets. We also use bootstrapping described in Section 2.3 to further enhance the unsupervised learning results.

Supervised learning. To compare how good dataless classification can be, we implemented supervised baselines for both data sets. We use simple BOW representation of documents (tokenized with stop words removed by Lucene), and use Liblinear [Fan *et al.*, 2008] as the classifier. Particularly, we use the L2-regularized and L2-loss linear support vector classification for all the experiments.

Cross-lingual word embedding. Another approach to perform cross-lingual dataless classification is to embed the words in both languages into the same semantic space, and then compare documents and labels in different languages in the same space. We use the compositional vector model (CVM) [Hermann and Blunsom, 2014] to generate our bi-lingual word embedding in a shared semantic space. CVM needs parallel corpora to train embedding for both languages. Following [Hermann and Blunsom, 2014], we train the models based on TED[7] and Europarl[8] data sets. TED data is derived from a spoken language translation data set[9]. It contains 13 languages from the TED talk transcriptions and their translations. Europarl data set is a popularly used parallel corpus for machine translation [Koehn, 2005]. It has 21 European languages that can be translated into English and vise versa. In this experiment, we select the ten relevant languages to the cross-lingual dataless classification tasks.

[5] https://github.com/mouuff/Google-Translate-API

[6] Google translate only supports 88 out of the 179 languages that our CLESA method can deal with.

[7] http://www.clg.ox.ac.uk/tedcldc/

[8] http://www.statmt.org/europarl/

[9] https://wit3.fbk.eu/

Table 3: Comparison on TED data set (averaged macro-F1 scores over 15 labels). Dataless naive (200): merging training and test data, selecting the top highest similarity scores as positive, and evaluating the F1 score on test set. Dataless bootstrapping: bootstrapping over the naive method. Dataless tuned: tuning a threshold on the training set, and applying it on the test data. "Average" excludes English.

	Supervised					Dataless (ESA)			Embed. (Tuned)	
	Full	10% mean	# training	15% mean	# training	Naive (200)	Bootstrapping	Tuned	TED	Europarl
English	0.508	0.316	104.7	0.360	157.1	0.389	0.405	0.440	–	–
Arabic	0.468	0.223	106.6	0.286	159.9	0.273	0.299	0.266	0.240	–
German	0.449	0.234	100.2	0.278	150.3	0.222	0.245	0.248	0.219	0.115
Spanish	0.525	0.303	106.1	0.331	159.2	0.289	0.301	0.293	0.245	0.163
French	0.547	0.353	104.9	0.324	157.4	0.205	0.228	0.206	0.253	0.157
Italian	0.535	0.294	104.2	0.315	156.3	0.191	0.197	0.226	0.289	0.177
Dutch	0.494	0.308	100.6	0.319	150.9	0.340	0.360	0.390	0.285	0.157
Polish	0.420	0.209	100.0	0.296	150.0	0.227	0.253	0.286	0.278	0.174
Pt-Br	0.502	0.271	100.3	0.296	150.5	0.307	0.331	0.287	0.250	0.171
Roman.	0.491	0.295	107.1	0.257	160.7	0.170	0.194	0.241	0.232	0.213
Russian	0.475	0.216	93.7	0.278	140.6	0.199	0.195	0.205	0.127	–
Turkish	0.426	0.176	95.2	0.252	142.8	0.333	0.354	0.395	0.248	–
Chinese	0.235	0.158	100.4	0.167	150.6	0.173	0.182	0.239	0.197	–
Average	0.468	0.258	101.8	0.289	152.8	0.255	0.273	0.286	0.238	0.166

We trained the CVM model using the parallel corpora with the default setting as well as the settings indicated in the paper [Hermann and Blunsom, 2014] using their software[10]. The length of the word vector was set to 128, the number of iterations was set to five, and the number of mini-batches was set to ten. We used the "additive" model with single mode (only using pairwise languages) and used the "doc-train" model to train on TED data and the "dbltrain" model to train on Europarl data.

4.2 TED Data Classification

The TED data set is a multi-label classification data set containing 15 labels of topics which are extracted from the most frequent keywords in the data set. The data has already been organized into subsets according to each label. Thus, we treat the problem as a binary classification for each label. Since the data is imbalanced, we find that training a supervised binary classifier and using the default threshold to determine which one is positive is not effective enough. Thus, we randomly split the provided training set into 70% training and 30% validation sets. Then we use the training set to train a model and use the validation set to tune the threshold. We average the results over ten trials to select the best threshold. Then we train a new model using the full training data and apply the new model and the tuned threshold to the test set. Besides using the full training set, we also use 10% and 15% of the full training set to do the same supervised procedure, respectively. We report the averaged F1 scores over 10 trials for 15 labels with supervised learning in Table 3. The fully supervised learning results are comparable with the best results shown in Table 4 in [Hermann and Blunsom, 2014].

For dataless classification, since there is only one label for each binary classification problem, it is only possible to use one similarity to select the most similar documents and label them as positive. Therefore, we perform a naive dataless classification as follows. First, we merge the training and testing data sets, which contains around 1200 documents for each label and each language. Then we select the 200 highest similarities between each label and the documents, and label them as positive. For bootstrapping, we initially label 50 positive and 500 negative examples respectively, and train a classifier, and then iteratively label 5 positive and 50 negative more documents in each bootstrapping step. We also combine the bootstrapping results with the top 200 positive documents labeled with pure dataless classification to ensure good recall. In addition, we also use another setting to verify the cross-lingual ESA similarity. We use the training set to tune a threshold for the similarities computed by cross-lingual ESA between both labels and documents. Then we apply the threshold to the test set to classify the documents. We call this the "tuned dataless classification."

From Table 3 we can see that naive dataless classification with the top 200 documents performs worst among the three settings, while bootstrapping is in the middle and tuned dataless classification performs the best. Compared to supervised learning, dataless classification is comparable to supervised learning with 10% labeled data, and a little worse than supervised learning with 15% labeled data. This result is consistent with the results shown in the original monolingual dataless classification [Chang *et al.*, 2008; Song and Roth, 2014]. It is amazing that for Chinese document classification, dataless classification is even better than the fully supervised learning. This may be because that Chinese typically uses fewer segmented words than English to represent the same meanings (0.68 million tokens in Chinese vs. 2.99 million tokens in English in TED). Then when classification is conducted on BOW features, there are fewer overlapped words among documents in Chinese, as compared to English and other languages.

We also use the multi-lingual embedding results of CVM for the dataless setting. We only show the results based on the tuned classification approach (tuning threshold based on training and applying the threshold for testing) in Table 3. Since Europarl data cannot cover all the language pairs used in TED, we only report the ones that it can cover. From the results we can see that even though the Europarl data set is much larger than TED, the embedding results trained based on TED data are much better than the embedding trained

[10] https://github.com/karlmoritz/CVM

Table 4: Comparison on RCV1/RCV2 data sets (top level, four categories). S.400: supervised learning with 400 training data. S.800: supervised learning with 800 training data. Datal.: dataless. Boots.: dataless with bootstrapping. E.(T.): word embedding using CVM trained on TED data, document embedding with average word embedding. E.(E.): word embedding using CVM trained on Europarl data. "Average" excludes English.

		micro-F1					
	#Doc.	S.400	S.800	Datal.	Boots.	E.(T.)	E.(E.)
RCV1	23,149	0.691	0.786	0.653	0.742	–	–
Danish	11,185	0.589	0.630	0.317	0.364	–	0.352
German	116,212	0.424	0.492	0.613	0.724	0.396	0.305
Spanish	18,655	0.645	0.651	0.647	0.667	0.156	0.290
Sp.-latam	79,775	0.241	0.250	0.644	0.554	0.376	0.536
French	85,393	0.307	0.467	0.653	0.762	0.578	0.334
Italian	28,406	0.553	0.607	0.528	0.542	0.323	0.274
Japanese	65,499	0.548	0.595	0.324	0.534	–	–
Dutch	1,794	0.140	0.160	0.387	0.395	0.125	0.205
Norwegian	9,409	0.510	0.564	0.252	0.329	–	–
Portuguese	8,841	0.546	0.613	0.428	0.375	0.101	0.257
Russian	17,487	0.499	0.523	0.309	0.418	0.334	0.323
Swedish	15,732	0.454	0.518	0.466	0.618	–	0.330
Chinese	28,964	0.672	0.723	0.537	0.690	0.241	–
Average		0.487	0.541	0.470	0.536	0.292	0.320

based on the Europarl data set. The dataless classification based on TED embedding is also worse than cross-lingual ESA. This is also reasonable since (1) Wikipedia contains more data for most of the languages; (2) ESA representation uses more global context in a document while embedding methods consider context more locally. For topical level judgment, ESA may be a better choice to represent the semantic meaning. We also verified this conclusion by testing English language with word2vec [Mikolov *et al.*, 2013b; 2013a] trained with Skipgram model, vector length as 128, and window size as five on the whole English Wikipedia. The tuned dataless classification result is 0.346, which is less than the English ESA (0.440) shown in Table 3. Similar results have also been shown in previous monolingual dataless classification [Song and Roth, 2014].

4.3 RCV2 Data Classification

We conducted similar experiments on the RCV2 data set, which is a multi-lingual extension of RCV1 [Lewis *et al.*, 2004]. We use the training set split by [Lewis *et al.*, 2004] but with the original documents instead of the ones after stemming. Same as RCV1 data, RCV2 is newswire stories from Reuters Ltd under the Factiva news category taxonomies. There are different categorization methods in RCV1, e.g., topical or regional. For the topical categories, it contains 103 categories including all nodes except for root in the hierarchy. The maximum depth is four, and 82 nodes are leaves. Following cross-lingual document classification [Klementiev *et al.*, 2012], we use the top level categories for evaluation. There are four categories which are GCAT (government social), ECAT (economics), MCAT (markets), and CCAT (corporate industrial). We aggregate all the subtree's English descriptions for each category as the category description. RCV2 data contains documents in 13 languages. The statistics of the document numbers are shown in Table 4.

We use the linear classifier trained on BOW as the baseline method. We train the classifiers with 400 and 800 randomly selected examples for each language respectively. We report the average over 10 trials for supervised learning results. For dataless classification, we have four classes and we choose the best label for each document based on the highest similarity between a document and the label descriptions. Then for bootstrapping, we use the standard procedure to initialize 100 documents for each class using pure similarity based dataless classification, and then iteratively label 100 more documents for each class and stop after three iterations. From Table 4 we can see that dataless classification with bootstrapping is comparable to supervised learning using between 400 and 800 labeled documents.

For the dataless classification based on multi-lingual embedding, we can see that the embedding trained on TED performs worse than embedding on Europarl. Compared to the TED classification results where TED embedding is much better, now both TED embedding and Europarl embedding are applied to out-of-domain examples (RCV2 words). Thus, when changing the domain, the size of the training corpus matters. Again, cross-lingual ESA is better than cross-lingual embedding. This is reasonable, since for embedding, we average all the word vectors to represent a document and a label. Thus some information may be lost. In the CVM paper, the authors also verified that embedding methods are worse than the original BOW representation for supervised learning methods [Hermann and Blunsom, 2014]. Moreover, the dataless classification with word2vec embedding [Mikolov *et al.*, 2013a] with 128 dimensions trained on English Wikipedia for RCV1 is 0.561. This again verifies that embedding currently under-performs ESA for dataless classification.

5 Conclusion

This paper shows that it is possible to classify documents in multiple languages into an English label space, within the dataless framework, without any training data. We propose to use cross-lingual ESA as the cross-lingual text representation into which we map the target documents and the label space. The experiments conducted on 88 languages derived from 20-newsgroups data show that for 28 languages, the pure dataless classification can achieve greater than 0.8 accuracy. We also tested on two multi-lingual benchmark data sets, i.e., TED and RCV2 data sets, showing that dataless classification is comparable to supervised learning with about 100 labeled documents per label. An important future direction would be to formulate an approach to enable dataless classification for lower-resource languages such as Uzbek and Hausa.

Acknowledgment

This work was supported by DARPA under agreement numbers HR0011-15-2-0025 and FA8750-13-2-0008; and by the U.S. Department of Homeland Security under Number 2009-ST-061-CCI002-07. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the organizations that supported the work.

References

[Amini and Goutte, 2010] Massih-Reza Amini and Cyril Goutte. A co-classification approach to learning from multilingual corpora. *Machine Learning*, 79(1-2):105–121, 2010.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Brown *et al.*, 1992] Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[Chang *et al.*, 2008] Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *AAAI*, pages 830–835, 2008.

[Chapelle *et al.*, 2006] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

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

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Gabrilovich and Markovitch, 2009] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34(1):443–498, 2009.

[Hermann and Blunsom, 2014] Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. In *ACL*, pages 58–68, 2014.

[Klementiev *et al.*, 2012] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474, 2012.

[Koehn, 2005] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit*, pages 79–86, 2005.

[Lang, 1995] Ken Lang. Newsweeder: Learning to filter netnews. In *ICML*, pages 331–339, 1995.

[Lewis *et al.*, 2004] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, December 2004.

[Liang, 2005] Percy Liang. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology, 2005.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119. 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *NAACL-HLT*, pages 746–751, 2013.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[Potthast *et al.*, 2008] Martin Potthast, Benno Stein, and Maik Anderka. A wikipedia-based multilingual retrieval model. In *ECIR*, pages 522–530, 2008.

[Prettenhofer and Stein, 2010] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *ACL*, pages 1118–1127, 2010.

[Shi *et al.*, 2010] Lei Shi, Rada Mihalcea, and Mingjun Tian. Cross language text classification by model translation and semi-supervised learning. In *EMNLP*, pages 1057–1067, 2010.

[Shirakawa *et al.*, 2014] Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. MLJ: Language-independent real-time search of tweets reported by media outlets and journalists. *Proceedings of the VLDB Endowment*, 7(13):1605–1608, 2014.

[Song and Roth, 2014] Yangqiu Song and Dan Roth. On dataless hierarchical text classification. In *AAAI*, pages 1579–1585, 2014.

[Song and Roth, 2015] Yangqiu Song and Dan Roth. Unsupervised sparse vector densification for short text similarity. In *NAACL-HLT*, pages 1275–1280, 2015.

[Sorg and Cimiano, 2012] Philipp Sorg and Philipp Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data and Knowledge Engineering*, 74:26–45, 2012.

[Turian *et al.*, 2010] Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384–394, 2010.

[Xiao and Guo, 2013] Min Xiao and Yuhong Guo. Semi-supervised representation learning for cross-lingual text classification. In *EMNLP*, pages 1465–1475, 2013.