Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-Language Text Classification

Aditya Mogadala
Institute AIFB
Karlsruhe Institute of Technology
Karlsruhe, Germany
aditya.mogadala@kit.edu

Achim Rettinger
Institute AIFB
Karlsruhe Institute of Technology
Karlsruhe, Germany
rettinger@kit.edu

Abstract

In many languages, sparse availability of resources causes numerous challenges for textual analysis tasks. Text classification is one of such standard tasks that is hindered due to limited availability of label information in low-resource languages. Transferring knowledge (i.e. label information) from high-resource to low-resource languages might improve text classification as compared to the other approaches like machine translation. We introduce BRAVE (*Bilingual paRAgraph VEctors*), a model to learn bilingual distributed representations (i.e. embeddings) of words without word alignments either from sentence-aligned parallel or label-aligned non-parallel document corpora to support cross-language text classification. Empirical analysis shows that classification models trained with our bilingual embeddings outperforms other state-of-the-art systems on three different cross-language text classification tasks.

1 Introduction

The availability of language-specific annotated resources is crucial for the efficiency of natural language processing tasks. Still, many languages lack rich annotated resources that support various tasks such as part-of-speech tagging, dependency parsing and text classification. While the growth of multilingual information on the web has provided an opportunity to build these missing annotated resources, but still lots of manual effort is required to achieve high quality resources for every language separately.

Another possibility is to utilize the unlabeled data present in those languages or transfer knowledge from annotation-rich languages. For the first alternative, recent advancements made in learning monolingual distributed representations of words (Mikolov et al., 2013a; Pennington et al., 2014; Levy and Goldberg, 2014) (i.e. monolingual word embeddings) capturing syntactic and semantic information in an unsupervised manner was useful in numerous NLP tasks (Collobert et al., 2011). However, this may not be sufficient for several other tasks such as cross-language information retrieval (Grefenstette, 2012), cross-language word semantic similarity (Vulić and Moens, 2014), cross-language text classification (CLTC, henceforth) (Klementiev et al., 2012; Xiao and Guo, 2013; Prettenhofer and Stein, 2010; Tang and Wan, 2014) and machine translation (Zhao et al., 2015) due to irregularities across languages. In these kind of scenarios, transfer of knowledge can be useful.

Several approaches (Hermann and Blunsom, 2014; Sarath Chandar et al., 2014; Gouws et al., 2015; Coulmance et al., 2015) tried to induce monolingual distributed representations into a language independent space (i.e. bilingual or multilingual word embeddings) by jointly training on pair of languages. Although the overall goal of these approaches is to capture linguistic regularities in words that share same semantic and syntactic space across languages, they differ in their implementation. One set of methods either perform offline alignment of trained monolingual embeddings or jointly-train both monolingual and cross-lingual objectives, while the other set uses only cross-lingual objective. Jointly-trained or offline alignment methods can be further divided based on the type of par-

Proceedings of NAACL-HLT 2016, pages 692–702,
San Diego, California, June 12-17, 2016.

Cross-Language Setups			
Objective	Method	Tasks	Parallel Corpus
Monolingual+ Cross-lingual	(Klementiev et al., 2012)	CLDC	Word-Aligned
	(Zou et al., 2013)	MT,NER	Word-Aligned
	(Mikolov et al., 2013b)	MT	Word-Aligned
	(Faruqui and Dyer, 2014)	Word Similarity	Word-Aligned
	(Lu et al., 2015)	Word Similarity	Word-Aligned
	(Gouws and Søgaard, 2015)	POS,SuS	Word-Aligned
	(Gouws et al., 2015)	CLDC,MT	Sentence-Aligned
	(Coulmance et al., 2015)	CLDC,MT	Sentence-Aligned
Cross-lingual	(Hermann and Blunsom, 2014)	CLDC	Sentence-Aligned
	(Sarath Chandar et al., 2014)	CLDC	Sentence-Aligned
	(Luong et al., 2015)	Word Similarity, CLDC	Sentence-Aligned
	(Pham et al., 2015)	CLDC	Sentence-Aligned

Table 1: Summary of bilingual or multilingual embedding methods that support Cross-language Document Classification (CLDC), Machine Translation (MT), Named Entity Recognition (NER), Part-of-Speech Tagging (POS), Super Sense Tagging (SuS).

allel corpus (e.g. word-aligned, sentence-aligned) they use for learning the cross-lingual objective. Table 1 summarizes different setups to learn bilingual or multilingual embeddings for the various tasks.

Methods in the Table 1 that use word-aligned parallel corpus as offline alignment (Mikolov et al., 2013b; Faruqui and Dyer, 2014) assume single correspondence between the words across languages and ignore polysemy. While, the jointly-train methods (Klementiev et al., 2012) that use word-alignment parallel corpus and consider polysemy perform computationally expensive operation of considering all possible interactions between the pairs of words in vocabulary of two different languages. Methods (Hermann and Blunsom, 2014; Sarath Chandar et al., 2014) that overcame the complexity issues of word-aligned models by using sentence-aligned parallel corpora limits themselves to only cross-lingual objective, thus making these approaches unable to explore monolingual corpora. Jointly-trained models (Gouws et al., 2015; Coulmance et al., 2015) overcame the issues of both word-aligned and purely cross-lingual objective models by using monolingual and sentence-aligned parallel corpora. Nonetheless, these approaches still have certain drawbacks such as usage of only bag-of-words from the parallel sentences ignoring order of words. Thus, they are missing to capture the non-compositional meaning of entire sentence. Also, learned bilingual embeddings were heavily biased towards the sampled sentence-aligned parallel corpora. It is also sometimes hard to acquire sentence-level parallel corpora for every language pair. To subdue this concern, few approaches (Rajendran et al., 2015) used pivot languages like English or comparable document-aligned corpora (Vulić and Moens, 2015) to learn bilingual embeddings specific to only one task.

This major downside can be observed in other aforementioned methods also, which are inflexible to handle different types of parallel corpora and have a tight-binding between cross-lingual objectives and the parallel corpora. For example, a method using sentence-level parallel corpora cannot be altered to leverage document-level parallel corpora (if available) that might have better performance for some tasks. Also, none of the approaches do leverage widely available label/class-aligned non-parallel documents (e.g. sentiment labels, multi-class datasets) across languages which share special semantics such as sentiment or correlation between concepts as opposed to parallel texts.

In this paper, we introduce BRAVE a jointly-trained flexible model that learns bilingual embeddings based on the availability of the type of corpora (e.g. sentence-aligned parallel or label/class-aligned non-parallel document) by just altering the cross-lingual objective. BRAVE leverages paragraph vector embeddings (Le and Mikolov, 2014) of the monolingual corpora to effectively conceal semantics of the text sequences across languages and build a cross-lingual objective. Method closely related to our approach is by Pham et al. (2015) who uses shared context sentence vector across lan-

guages to learn multilingual text sequences.

The main contributions of this paper are:

- We jointly train monolingual part of parallel corpora with the improved cross-lingual alignment function that extends beyond bag-of-word models.
- Introduced a novel approach to leverage non-parallel data sets such as label or class aligned documents in different languages for learning bilingual cues.
- Experimental evaluation on three different CLTC tasks, namely cross-language document classification, multi-label classification and cross-language sentiment classification using learned bilingual word embeddings.

2 Related Work

Most of the related work can be associated to the approaches that aim to learn latent topics across languages or distributed representations of the words and larger pieces of text for supporting various cross-lingual tasks.

2.1 Cross-Language Latent Topics

Various approaches have been proposed to identify latent topics in monolingual (Blei, 2012; Rus et al., 2013) and multilingual (Mimno et al., 2009; Fukumasu et al., 2012) scenarios for cross-language semantic word similarity and document comparison. Extraction of cross-language latent topics or concepts use context-insensitive (Zhang et al., 2010) and context-sensitive methods (Vulić and Moens, 2014) to build word co-occurrence statistics for document representations.

2.2 Distributed Representations

Continuous word representations (Bengio et al., 2003; Mikolov et al., 2013a; Pennington et al., 2014) was further extended to multilingual (Hermann and Blunsom, 2014; Kočiský et al., 2014; Coulmance et al., 2015), bilingual (Gouws et al., 2015; Vulić and Moens, 2015; Luong et al., 2015) and polylingual (Al-Rfou et al., 2013) settings by projecting multiple or pair of languages into the shared semantic space. Also, word representations were extended to meet larger textual units like phrases, sentences and documents either monolingual (Socher et al., 2012; Le and Mikolov, 2014) or bilingual (Pham et al., 2015). Some approaches fine tuned the embeddings for specific tasks such as cross-lingual sentiment analysis (Zhou et al., 2015b), cross-language POS tagging (Gouws and Søgaard, 2015), machine translation (Cho et al., 2014) etc.

3 BRAVE Model

In this section, we present the BRAVE model along with its variations whose aim is to learn bilingual embeddings that can generalize across different languages.

3.1 Bilingual Paragraph Vectors (BRAVE)

Most of the NLP tasks require fixed-length vectors. Tasks like CLTC also require fixed-length vectors to incorporate inherent semantics of sentences or documents. Distributed representation of sentences and documents i.e. paragraph vectors (Le and Mikolov, 2014) are designed to out-perform certain text classification tasks by overcoming constraints posed by the bag-of-words models.

Here, we leverage paragraph vectors distributed memory model (PV-DM) as the monolingual objective $\mathcal{M}(\cdot)$ and jointly optimize with bilingual regularization function $\varphi(\cdot)$ for learning bilingual embeddings similar to the earlier approaches (Gouws et al., 2015; Coulmance et al., 2015). Equation 1 shows the formulation of the overall objective function that is minimized.

$$\mathcal{L} = \min_{\theta^{l_1}, \theta^{l_2}} \sum_{l \epsilon \{l_1, l_2\}} \sum_{C^l} \mathcal{M}^l(w_t, h; \theta^l) + \frac{\lambda \varphi(\theta^{l_1}, \theta^{l_2})}{2} \tag{1}$$

Here, C^l represent the corpus of individual languages (i.e. l_1 or l_2). Given any sequence of words $(w_1^l, w_2^l ... w_T^l)$ in C^l, w_t is the predicted word in a context h constrained on paragraph p (i.e. sentence or document) and sequence of words.

Formally, the first term (i.e. $\mathcal{M}(\cdot)$) in the Equation 1 maximizes the average log probability based on word vector matrix W^l and a unique paragraph vector matrix P^l. Equation 2 represents the average log probability.

$$\mathcal{M}^l(w_t, h; \theta^l) = \frac{\Sigma_{t=k}^{T-k} y_{w_t}^l - log(\sum_i e^{y_i^l})}{T} \tag{2}$$

where each y_i^l is log-probability of predicted word i and is given by Equation 3.

$$y^l = b + Uh(w_{t-k}^l ....w_{t+k}^l; W^l, P^l) \quad (3)$$

To optimize for efficiency, hierarchical softmax (Mnih and Hinton, 2009) is used in training with U and b as parameters. Binary Huffmann tree is utilized to represent hierarchial softmax (Mikolov et al., 2013a). Analogous to Pham et al., (2015), we also derive h by concatenating paragraph vector from P^l with the average of word vectors in W^l. This helps to fine tune both word and paragraph vectors independently.

Now, to capture the bilingual cues, the regularization function ($\varphi(\cdot)$) is learned in two different ways. In the first approach a sentence-aligned parallel corpora is used, while in the second approach a label-aligned document corpora.

3.2 BRAVE with Sentence-Aligned Parallel corpora (BRAVE-S)

To compute the bilingual regularization function $\varphi(\cdot)$, we slightly deviate from earlier approaches (Gouws et al., 2015). Instead of simply performing L_2-loss between the mean of word vectors in each sentence pair ($s_j^{l_1}$,$s_j^{l_2}$) of the sentence-aligned parallel corpus (PC) at each training step. We use the concept of elastic net regularization (Zou and Hastie, 2005) and employ linear combination of L_2-loss between *sentence paragraph vectors* $SP_j^{l_1}$ and $SP_j^{l_2} \in R^d$ precomputed from the monolingual term $\mathcal{M}(\cdot)$ with L_2-loss between the mean of word vectors observed in sentences. This induces a constraint on the usage of monolingual part of parallel training data to learn $\mathcal{M}(\cdot)$. At the same time, it has an advantage of using combination of paragraph and word vectors which combines compositional and non-compositional meanings of sentences.

Also, it eliminates the need for word-alignment and makes an assumption that each word observed in the sentence of language l_1 can potentially find its alignment in the sentence of language l_2. Theoretically, low value of $\varphi(\cdot)$ ensures that words across languages which are similar are embedded closer to each other. Equation 4 shows the regularization term.

$$\alpha||SP_j^{l_1} - SP_j^{l_2}||^2 + (1-\alpha)||\frac{1}{m}\sum_{w_i \epsilon s_j^{l_1}}^{m} W_i^{l_1} - \frac{1}{n}\sum_{w_k \epsilon s_j^{l_2}}^{n} W_k^{l_2}||^2 \quad (4)$$

Where $W_i^{l_1}$ and $W_k^{l_2}$ represent word embeddings obtained for the words w_i and w_k in each sentence (s_j) of length m and n in languages l_1 and l_2 respectively.

3.3 BRAVE with Non-Parallel Document Corpora (BRAVE-D)

Sometimes it is hard to acquire sentence-aligned parallel corpora for many languages. Availability of non-parallel corpora such as topic-aligned (e.g. Wikipedia) or label/class-aligned document corpora (e.g. sentiment analysis and multi-class classification data sets) in different languages can be leveraged to learn bilingual embeddings for performing CLTC. Earlier approaches like CL-LSI (Dumais et al., 1997) and CL-KCCA (Vinokourov et al., 2003) were used to learn bilingual document spaces for the tasks comparable to CLTC. Although these approaches provide decent results, they face serious scalability issues and are mostly limited to Wikipedia. Cross-lingual latent topic extraction models (Vulić and Moens, 2014) showed promising results for the tasks like word-level or phrase-level translations, but have certain drawbacks for CLTC tasks.

Here, we propose a two step approach to build bilingual embeddings with label/class-aligned document corpora.

- In the first step, we perform manifold alignment using Procrustes analysis (Wang and Mahadevan, 2008) between sets of documents belonging to same class/label in different languages. This will help to identify the closest alignment of a document in language l_1 with a document in another language l_2.

- In the second step, we use the pair of partially aligned documents belonging to same class or label in different languages to extract bilingual cues similar to the approach mentioned in § 3.2. Only difference being paragraph vector is learned for the entire document.

Step-1:
Let S^{l_1} and S^{l_2} be the sets containing languages l_1 and l_2 training documents associated to label or a class. Below, we provide the three step procedure to attain partial alignment between the documents present in these sets.

- Learning low-dimensional embeddings of the sets (S^{l_1}, S^{l_2}) is key for alignment. We use **document paragraph vectors** (Le and Mikolov, 2014) to learn low-dimensional embeddings of the documents in each language. Let X^{l_1} and X^{l_2} be the low-dimensional embeddings of S^{l_1} and S^{l_2} respectively.

- To find the optimal values of transformation, Procrustes superimposition is done by translating, rotating and scaling the objects (i.e. rows of X^{l_2} is transformed to make it similar to the rows of X^{l_1}). Transformation is achieved by

  - **Translation:** Taking mean of all the members of set to make centroids $(\sum_{i=1}^{|S^{l_1}|} \frac{X^{l_1}}{|S^{l_1}|}, \sum_{i=1}^{|S^{l_2}|} \frac{X^{l_2}}{|S^{l_2}|})$ lie at origin.
  - **Scaling and Rotation:** The rotation and scaling that maximizes the alignment is given by orthogonal matrix (Q) and scaling factor (k). They are obtained by minimizing orthogonal Procrustes problem (Schönemann, 1966) and is provided by Equation 5.

$$\arg\min_{k,Q} ||X^{l_1} - X_*^{l_2}||_F \qquad (5)$$

    where $X_*^{l_2}$ a matrix of transformed X^{l_2} values given by $kX^{l_2}Q$ and $||.||_F$ is the Frobenius norm constrained over $Q^TQ = I$.

- If $S_*^{l_2}$ represents the new document set obtained after identifying the close alignment among documents in S^{l_1} and S^{l_2} with cosine similarity between X^{l_1} and $X_*^{l_2}$, then the partially aligned corpora $\{S^{l_1}, S_*^{l_2}\}$ contains one-to-one correspondence between the two languages documents that are used to learn bilingual cues in the second step.

From perturbation theory of spectral spaces (Kostrykin et al., 2003) it can be understood that the difference between low-dimensional embedding subspaces (i.e. X^{l_1} and $X_*^{l_2}$) is always bounded, thus the new alignment obtained between document sets $\{S^{l_1}, S_*^{l_2}\}$ is insensitive to perturbations. Which also means that Procrustes analysis has provided best possible document alignments.

Step-2:
Now, document pairs ($d_j^{l_1}$,$d_j^{l_2}$) of the partially-aligned corpus (PAC) is used to compute bilingual regularization function $\varphi(\cdot)$. At each training step, L_2-loss of precomputed *document paragraph vectors* DP_j^{l1} and $DP_j^{l2} \in R^d$ obtained from the monolingual term $\mathcal{M}(\cdot)$ is combined with the L_2-loss between vector of words weighted by the probability of their occurrence in a particular label/class of entire **PAC**. Consideration of word probabilities will help to induce label/class specific information. Equation 6 provides the regularization term.

$$\alpha||DP_j^{l_1} - DP_j^{l_2}||^2 + (1-\alpha)||\sum_{w_i \epsilon d_j^{l_1}}^{m} \frac{p_{w_i} W_i^{l_1}}{\sum_m p_{w_i}} - \sum_{w_k \epsilon d_j^{l_2}}^{n} \frac{q_{w_k} W_k^{l_2}}{\sum_n q_{w_k}}||^2 \qquad (6)$$

Where w_i,w_k are words and their embeddings $W_i^{l_1}$,$W_k^{l_2}$ observed in each document (d_j) of length m and n in languages l_1 and l_2 respectively. While, p_{w_i} and q_{w_k} represents probability of occurrence of words w_i and w_k in a specific label/class of entire **PAC**. Figure- 1 shows overall goal of both the approaches.

4 Experiments

In this section, we report results on three different CLTC tasks to comprehend whether our learned bilingual embeddings are semantically useful across languages. First, cross-language document classification (CLDC) task proposed by Klementiev et al. (2012) using the subset of Reuters RCV1/RCV2 corpora (Lewis et al., 2004). Second, a multi-label CLDC task with more languages using TED corpus[1] of Hermann et al. (2014) . Subsequently, a cross-language sentiment classification (CLSC) proposed by Prettenhofer et al., (2010) on a multi-domain sentiment dataset.

[1] http://www.clg.ox.ac.uk/tedcorpus

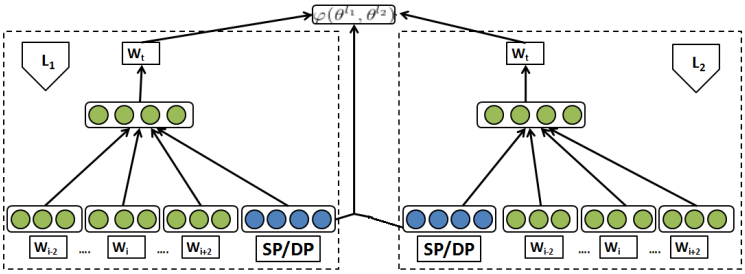

Figure 1: Bilingual word embeddings learned using sentence or document paragraph vectors (SP/DP) along with word vectors.

4.1 Parallel and Non-Parallel Corpora

For sentence-aligned parallel corpora, Europarl-v7 [2](EP) is used as both monolingual and parallel training data. While for label-aligned non-parallel document corpora, only training and testing collections of the cross-language multi-domain Amazon product reviews(CL-APR) (Prettenhofer and Stein, 2010) corpus with sentiment labels is used.

4.2 Implementation

Our implementation launches monolingual paragraph vector (Le and Mikolov, 2014) threads for each language along with bilingual regularization thread. Word and paragraph embeddings matrices are initialized with normal distribution ($\mu = 0$ and $\sigma^2 = 0.1$) for each language and all threads access them asynchronously. Following Pham et al. (2015) suggested combination (P=5*W) of paragraph and word embeddings, we chose paragraph embeddings with dimensionality of 200 and 640 when word embeddings are of 40 and 128 dimensions respectively. Asynchronous stochastic gradient descent (ASGD) is used to update parameters (i.e. P^l,W^l,U and b) and train the model.

For each training pair in parallel or non-parallel corpora, initially monolingual threads sample context h with window size of 8 from a random paragraph (i.e. sentence or document) in each language. Then the bilingual regularization thread along with monolingual threads make update to parameters asynchronously. Learning rate is set to 0.001 which decrease with the increase of epochs, while α is chosen to be 0.6 (can be fine tuned based on empirical analysis) to give more weight to paragraph vectors. All models are trained for 50 epochs.

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

4.3 Document Representation

Documents are represented with tf-idf weighted sum of embedding vectors of the words that are present in them.

4.4 Results

The experimental results for each of the CLTC tasks are presented separately.

4.4.1 Cross-language Document Classification (CLDC) - RCV1/RCV2

Goal of this task is to classify target language documents with the labeled examples from the source language. To achieve it, we used the subset of Reuters RCV1/RCV2 corpora as the training and evaluation sets and replicated the experimental setting of Klementiev et al. (2012). From the English, German, French and Spanish collection of the dataset, only those documents are selected which was labeled with a single topic (i.e. CCAT, ECAT, GCAT and MCAT). For the classification experiments, 1000 labeled documents from source language are selected to train a multi-class classifier using averaged perceptron (Freund and Schapire, 1999; Collins, 2002) and 5000 documents were used as the testing data.

English-German, English-French and English-Spanish portion of **EP** corpora (i.e. each with around 1.9M sentence-pairs) is used both as monolingual and parallel training data with **BRAVE-S** approach to build vocabulary of around 85k English, 144k German, 119k French and 118k Spanish. While training and testing collections belonging to all domains in English-German, English-French languages of **CL-APR** ((i.e. around 12,000 document-pairs)) was used both as monolingual and partially aligned data with **BRAVE-D** approach to build vocabulary of around 21k English, 22k German and

Model	Dim	en → de	de → en	en → fr	fr → en	en → es	es → en
Majority class	40	46.8	46.8	22.5	25.0	15.3	22.2
MT	40	68.1	67.4	76.3	71.1	52.0	58.4
I-Matrix (Klementiev et al., 2012)	40	77.6	71.1	74.5	61.9	31.3	63.0
BAE-cr (Sarath Chandar et al., 2014)	40	**91.8**	74.2	**84.6**	74.2	49.0	64.4
CVM-Add (Hermann and Blunsom, 2014)	40	86.4	74.7	-	-	-	-
DWA (Kočiský et al., 2014)	40	83.1	75.4	-	-	-	-
BilBOWA (Gouws et al., 2015)	40	86.5	75	-	-	-	-
UnsupAlign (Luong et al., 2015)	40	87.6	77.8	-	-	-	-
Trans-gram (Coulmance et al., 2015)	40	87.8	78.7	-	-	-	-
BRAVE-S$_{(EP)}$	40	88.1	**78.9**	79.2	**77.8**	**56.9**	**67.6**
BRAVE-D$_{(CL\text{-}APR)}$	40	69.4	67.9	64.1	56.5	-	-
CVM-BI (Hermann and Blunsom, 2014)	128	86.1	79.0	-	-	-	-
UnsupAlign (Luong et al., 2015)	128	88.9	77.4	-	-	-	-
BRAVE-S$_{(EP)}$	128	89.7	**80.1**	82.5	**79.5**	**60.2**	**70.4**
BRAVE-D$_{(CL\text{-}APR)}$	128	70.4	70.6	66.2	57.6	-	-

Table 2: CLDC Accuracy with 1000 labeled examples on RCV1/RCV2 Corpus. en/de, en/fr and en/es results of Majority class, MT, I-Matrix and BAE-cr are adopted from Sarath Chandar et al., (2014)

18k French. Further, documents in the training and testing data of RCV1/RCV2 corpora are represented as described in § 4.3 with the vocabulary built. Table 2 shows the comparison of our approaches with the existing systems.

4.4.2 Multi-label CLDC - TED Corpus

To understand the applicability of our approaches to wider range of languages[3] and class labels, we perform experiments with the subset of TED corpus (Hermann and Blunsom, 2014). Aim of this task is same as § 4.4.1, but experiments were conducted with larger variety of languages and class labels. TED Corpus contains English transcriptions and their sentence-aligned translations for 12 languages from the TED conference. Entire corpus is further classified into 15 topics (i.e. class labels) based on the most frequent keywords appearing in them.

To conduct our experiments, we follow the *single* mode setting of Hermann et al. (2014) (i.e. embeddings are learned only from a single language pair). Entire language pair (i.e. en→L2) training data of the TED corpus is used both as monolingual and parallel training data to learn bilingual word embeddings with dimensionality of 128 using **BRAVE-S** approach. Bilingual word embeddings of 128 dimensions learned with **EP** and **CL-APR** are also used for comparison. Documents in the training and testing data of TED corpus are represented as described in § 4.3 using each of these embeddings. A multi-class classifier using averaged perceptron is built using training documents in source language to be applied on target language testing data for predicting the class labels. Table 3 shows the cumulative F1-scores.

4.4.3 Cross-language Sentiment Classification (CLSC)

The objective of the third CLTC task is to identify sentiment polarity (e.g. positive or negative) of the data in target language by exploiting the labeled data in source language. We chose subset of publicly available Amazon product reviews (CL-APR) (Prettenhofer and Stein, 2010) dataset mainly English(E), German(G) and French(F) languages belonging to three different product categories (books(B), dvds(D) and music(M)) to conduct our experiments. For each language-category pair, corpus consists of training, testing sets comprising 1000 positive and 1000 negative reviews each with an additional unlabeled reviews varying from 9,000 to 170,000.

We constructed 12 different CLSC tasks using different languages (i.e. E,G and F) for three categories (i.e. B,D and M). For example, EFM refers English music reviews as source language and French music reviews as target language. Bilingual word embeddings with dimensionality of 128 learned with

[3]Our goal is not to evaluate shared multilingual semantic representation.

Method	de	es	fr	it	nl	pt	po	ro	ru	tr
en → L2										
MT-*Baseline*	0.465	**0.518**	**0.526**	**0.514**	**0.505**	0.470	0.445	**0.493**	0.432	0.409
DOC/ADD	0.424	0.383	0.476	0.485	0.264	0.354	0.402	0.418	0.448	0.452
DOC/BI	0.428	0.416	0.445	0.473	0.219	0.400	0.403	0.467	0.421	0.457
BRAVE-S$_{(TED)}$	**0.484**	0.436	0.456	0.507	0.328	**0.506**	**0.453**	0.488	**0.456**	**0.491**
BRAVE-S$_{(EP)}$	0.418	0.365	0.387	0.418	0.284	0.454	0.412	0.424	-	-
BRAVE-D$_{(CL-APR)}$	0.385	-	0.212	-	-	-	-	-	-	-
L2 → en										
MT-*Baseline*	0.469	0.486	0.358	**0.481**	**0.463**	0.374	**0.460**	**0.486**	0.404	0.441
DOC/ADD	0.476	0.422	0.464	0.461	0.251	0.338	0.400	0.407	0.471	0.435
DOC/BI	0.442	0.365	**0.479**	0.460	0.235	0.380	0.393	0.426	**0.467**	0.477
BRAVE-S$_{(TED)}$	**0.492**	**0.495**	0.465	0.475	0.384	**0.388**	0.442	0.464	0.457	**0.484**
BRAVE-S$_{(EP)}$	0.458	0.404	0.437	0.443	0.338	0.312	0.374	0.418	-	-
BRAVE-D$_{(CL-APR)}$	0.366	-	0.278	-	-	-	-	-	-	-

Table 3: Cumulative F1-scores on TED Corpus using training data in English language and evaluation on other languages (i.e. German (de), Spanish (es), French (fr), Italian (it), Dutch (nl), Portugese (pt), Polish (po), Romanian (ro), Russian (ru) and Turkish (tr)) and vice versa. MT-*Baseline*, DOC/ADD, DOC/BI represents single language pair of Hermann et al., (2014) as document features. Underline shows the best results amongst embedding models.

Cross-Language Sentiment Classification (en→L2 and Vice versa)							
Task	CL-SCL	CL-SSMC	CL-SLF	CL-DCI$_{100}$	BSE	BRAVE-S (EP)	BRAVE-D (CL-APR)
EFB	79.86±0.22	83.05±0.26	82.61±0.25	82.30	-	72.24±0.31	82.57±0.33
EFD	78.80±0.25	82.70±0.20	82.70±0.45	82.40	-	74.95±0.25	**82.90±0.35**
EFM	75.95±0.31	80.46±0.20	80.19±0.40	81.05	-	72.80±0.20	**80.70±0.45**
FEB	77.26±0.22	80.05±0.26	80.48±0.33	-	-	75.45±0.38	80.28±0.21
FED	76.57±0.20	79.40±0.28	78.76±0.38	-	-	73.75±0.26	**79.80±0.15**
FEM	76.76±0.25	78.82±0.17	79.18±0.33	-	-	73.66±0.17	78.56±0.33
EGB	77.77±0.28	81.88±0.42	79.91±0.47	81.40	80.27±0.50	75.95±0.16	**81.75±0.45**
EGD	79.93±0.23	82.25±0.20	81.86±0.31	79.95	77.16±0.30	78.30±0.42	81.56±0.26
EGM	73.95±0.30	81.30±0.20	79.59±0.42	83.30	77.98±0.51	75.95±0.33	81.20±0.17
GEB	77.85±0.27	79.06±0.23	78.61±0.34	-	-	72.25±0.20	**80.23±0.17**
GED	77.83±0.33	80.89±0.16	80.27±0.35	-	-	73.28±0.23	**80.78±0.20**
GEM	77.37±0.34	79.85±0.17	79.80±0.26	-	-	74.41±0.22	**79.77±0.36**

Table 4: Average classification accuracies and standard deviations for 12 CLSC tasks. Results of other baselines are adopted from CL-SCL (Prettenhofer and Stein, 2010), CL-SSMC (Xiao and Guo, 2002), CL-SLF (Zhou et al., 2015a), CL-DCI$_{100}$ (Esuli and Fernandez, 2015) and BSE (Tang and Wan, 2014)

BRAVE-S and **BRAVE-D** are used to represent each review as described in § 4.3. To have fair comparison with earlier approaches, sentiment classification model is then trained with libsvm[4] default parameter settings using source language training reviews[5] to classify target language test reviews. Table 4 shows the accuracy and standard deviation results after we randomly chose subset of target language testing documents and repeated the experiment for 10 times for all CLSC tasks.

5 Discussion

First CLTC task (i.e. CLDC) results presented in Table 2 shows that BRAVE-S was able to outperform most of the existing systems. Success of BRAVE-S can be attributed to its ability to incorporate both non-compositional and compositional meaning observed in entire sentence and the individual words respectively. Thus making it different from other models which use only bag-of-words (Gouws et al., 2015) or bi-grams (Hermann and Blunsom, 2014).

Similarly, second CLTC task (i.e. multi-label

[4]https://www.csie.ntu.edu.tw/ cjlin/libsvm/

[5]We do not use 100 labeled target language reviews in model training, as it was shown by earlier approaches that 100 labeled target language reviews does not have much impact.

Top-3 Nearest Neighbors (Euclidean Distance)			
English Words	Models	German	French
great	BRAVE-S	wachstum	éminent
		super	maintenus
		spielen	m'efforcerai
	BRAVE-D	schärfe	festival
		mögen	interressante
		kraftvolle	attachant
bored	BRAVE-S	boykottiert	ennuyé
		leere	précédera
		ausgehen	compromettent
	BRAVE-D	ableben	réserve
		lichtblick	intensité
		traurigen	consterné

Table 5: Nearest Neighbors for English Words in German and French.

CLDC) results presented in Table 3 shows that BRAVE-S learned with the training data of TED corpus outperformed *single mode* DOC/* embedding models (Hermann and Blunsom, 2014), BRAVE-S learned with **EP** and BRAVE-D. The BRAVE-S(TED) was able to capture better linguistic regularities across languages that is more specific to the corpus, than the general purpose bilingual embeddings learned with **EP**. Though in some cases, all our embedding models could not outperform machine translation baseline. This can be due to the asymmetry between languages induced by the language specific words which could not find its equivalents in English.

Also, it can be apprehended from the Table 2 and Table 3 that BRAVE-D results are not as expected. Though being a general approach like BRAVE-S which can capture both non-compositional and compositional meaning from larger pieces of texts, minimal overlap of vocabulary learned with BRAVE-D using cross-language sentiment label-aligned corpora with other domains (i.e. Reuters and TED) produce unfavorable results. Thus, we understand that the choice of label/class-aligned corpora is crucial.

Final CLTC task (i.e. CLSC) results presented in Table 4 shows that BRAVE-D outperforms other baseline approaches in most of the cases. As BRAVE-D learns bilingual word embeddings using **CL-APR**, it was able to inherently encompass sentiment label information effectively like earlier approaches (Tang and Wan, 2014; Zhou et al., 2015b) than the general purpose embeddings learned using BRAVE-S with **EP** and similar approaches (Meng et al., 2012). Thus making it more suitable for sentiment classification task. Also unlike CL-SSMC (Xiao and Guo, 2002) and CL-SLF (Zhou et al., 2015a), BRAVE-D is not highly parameter dependent where the results of the former approaches show big variance based on the parameter settings. To visualize the difference in embeddings learned with BRAVE-S and BRAVE-D, we selected sentiment words and identified cross-language nearest neighbors in Table 5. It can be observed that BRAVE-D was able to identify better sentiment (either positive or negative) word neighbors than BRAVE-S.

6 Conclusion and Future Work

In this paper, we presented an approach that leverages paragraph vectors to learn bilingual word embeddings with sentence-aligned parallel and label-aligned non-parallel corpora. Empirical analysis exhibited that embeddings learned from both of these types of corpora have shown good impact on CLTC tasks. In future, we aim to extend the approach to learn multilingual semantic spaces with more labels/classes.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

References

R. Al-Rfou, B. Perozzi, and S. Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *CoNLL*.

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research.*, 3:1137–1155.

D. M. Blei. 2012. Probabilistic topic models. *Communications of the ACM.*, 55(4):77–84.

K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

M. Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *ACL-EMNLP.*, pages 1–8.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research.*, 12:2493–2537.

J. Coulmance, J. M. Marty, G. Wenzek, and A. Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings reyes- mannde= reginait- femmefr. In *EMNLP.*

S. T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval.*

A. Esuli and A. M. Fernandez. 2015. Distributional correspondence indexing for cross-language text categorization. In *Advances in Information Retrieval.*, pages 104–109.

M. Faruqui and C. Dyer. 2014. Improving vector space word representations using multilingual correlation. In *ACL.*

Y. Freund and R. E. Schapire. 1999. Large margin classification using the perceptron algorithm. *The Journal of Machine Learning Research.*, 37(3):277–296.

K. Fukumasu, Koji Eguchi, and Eric P. Xing. 2012. Symmetric correspondence topic models for multilingual text analysis. In *NIPS*, pages 1295–1303.

S. Gouws and A. Søgaard. 2015. Simple task-specific bilingual word embeddings. In *NAACL-HLT.*, pages 1386–1390.

S. Gouws, Y. Bengio, and G. Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML.*

G. Grefenstette. 2012. Cross-language information retrieval. *Springer Science and Business Media*, 2.

K. M. Hermann and P. Blunsom. 2014. Multilingual models for compositional distributed semantics. In *ACL.*

A. Klementiev, I. Titov, and B. Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING.*

T. Kočiský, K. M. Hermann, and P. Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *ACL*, pages 224–229.

V. Kostrykin, K. Makarov, and A. Motovilov. 2003. On a subspace perturbation problem. *American Mathematical Society.*, pages 3469–3476.

Q. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196.

O. Levy and Y. Goldberg. 2014. Dependency based word embeddings. In *ACL.*, pages 302–308.

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

A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *NAACL-HLT.*

T. Luong, H. Pham, and C. D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *First Workshop on Vector Space Modeling for Natural Language Processing.*, pages 151–159.

X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang. 2012. Cross-lingual mixture model for sentiment classification. In *ACL.*, pages 572–581.

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

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

D. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*, pages 880–889.

A. Mnih and G. E. Hinton. 2009. A scalable hierarchical distributed language model. In *NIPS.*, pages 1081–1088.

J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP.*

H. Pham, M. T. Luong, and C. D. Manning. 2015. Learning distributed representations for multilingual text sequences. In *NAACL-HLT*, pages 88–94.

P. Prettenhofer and B. Stein. 2010. Cross-language text classification using structural correspondence learning. In *ACL.*, pages 1118–1127.

J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran. 2015. Bridge correlational neural networks for multilingual multimodal representation learning. In *arXiv preprint arXiv:1510.03519.*

V. Rus, M. C. Lintean, R. Banjade, N. B. Niraula, and D. Stefanescu. 2013. Semilar: The semantic similarity toolkit. In *ACL(Conference System Demonstrations)*, pages 163–168.

A. P. Sarath Chandar, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha. 2014. An autoencoder approach to learning bilingual word representations. In *NIPS.*, pages 1853–1861.

P. H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *The Journal of Psychometrika.*, 31(1):1–10.

R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP*, pages 1201–1211.

X. Tang and X. Wan. 2014. Learning bilingual embedding model for cross-language sentiment classification. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences.*, volume 2, pages 134–141.

A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. 2003. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS.*, pages 1497–1504.

I. Vulić and M. F. Moens. 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *EMNLP.*

I. Vulić and M. F. Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL.*

C. Wang and S. Mahadevan. 2008. Manifold alignment using procrustes analysis. In *ICML.*, pages 1120–1127.

M. Xiao and Y. Guo. 2002. Semi-supervised matrix completion for cross-lingual text classification. In *AAAI.*

M. Xiao and Y. Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *EMNLP.*, pages 1465–1475.

D. Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *ACL*, pages 1128–1137.

K. Zhao, H. Hassan, and M. Auli. 2015. Learning translation models from monolingual continuous representations. In *NAACL-HLT*.

G. Zhou, T. He, J. Zhao, and W. Wu. 2015a. A subspace learning framework for cross-lingual sentiment classification with partial parallel data. In *IJCAI.*

H. Zhou, L. Chen, F. Shi, and D. Huang. 2015b. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *ACL*, pages 430–440.

H. Zou and T. Hastie. 2005. Regularization and variable selection via the elastic net. *The Journal of the Royal Statistical Society: Series B (Statistical Methodology).*, 67(2):301–320.

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