

# Unsupervised Cross-lingual Transfer of Word Embedding Spaces

Ruochen Xu, Yiming Yang, Naoki Otani, Yuexin Wu

Carnegie Mellon University

{ruochenx, yiming, notani, yuexinw}@cs.cmu.edu

## Abstract

Cross-lingual transfer of word embeddings aims to establish the semantic mappings among words in different languages by learning the transformation functions over the corresponding word embedding spaces. Successfully solving this problem would benefit many downstream tasks such as to translate text classification models from resource-rich languages (e.g. English) to low-resource languages. Supervised methods for this problem rely on the availability of cross-lingual supervision, either using parallel corpora or bilingual lexicons as the labeled data for training, which may not be available for many low resource languages. This paper proposes an unsupervised learning approach that does not require any cross-lingual labeled data. Given two monolingual word embedding spaces for any language pair, our algorithm optimizes the transformation functions in both directions simultaneously based on distributional matching as well as minimizing the back-translation losses. We use a neural network implementation to calculate the Sinkhorn distance, a well-defined distributional similarity measure, and optimize our objective through back-propagation. Our evaluation on benchmark datasets for bilingual lexicon induction and cross-lingual word similarity prediction shows stronger or competitive performance of the proposed method compared to other state-of-the-art supervised and unsupervised baseline methods over many language pairs.

## 1 Introduction

Word embeddings are well known to capture meaningful representations of words based on large text corpora (Mikolov et al., 2013; Pennington et al., 2014). Training word vectors using monolingual corpora is a common practice in various NLP tasks. However, how to establish cross-lingual semantic mapping among mono-

lingual embeddings remain an open challenge as the availability of resources and benchmarks are highly imbalanced across languages.

Recently, increasing effort of research has been motivated to address this challenge. Successful cross-lingual word mapping will benefit many cross-lingual learning tasks, such as transforming text classification models trained in resource-rich languages to low-resource languages. Downstream applications include word alignment, text classification, named entity recognition, dependency parsing, POS-tagging, and more (Søgaard et al., 2015). Most methods for cross-lingual transfer of word embeddings are based on supervised or semi-supervised learning, i.e., they require cross-lingual supervision such as human-annotated bilingual lexicons and parallel corpora (Lu et al., 2015; Smith et al., 2017; Artetxe et al., 2016). Such a requirement may not be met for many language pairs in the real world.

This paper proposes an unsupervised approach to the cross-lingual transfer of monolingual word embeddings, which requires zero cross-lingual supervision. The key idea is to optimize the mapping in both directions for each language pair (say A and B), in the way that the word embedding translated from language A to language B will match the distribution of word embedding in language B. And when translated back from B to A, the word embedding after two steps of transfer will be maximally close to the original word embedding. A similar property holds for the other direction of the loop (from B to A and then from A back to B). Specifically, we use the Sinkhorn distance (Curi, 2013) to capture the distributional similarity between two set of embeddings after transformation, which we found empirically superior to the KL-divergence (Zhang et al., 2017a) and distance to nearest neighbor (Artetxe et al., 2017; Conneau et al., 2017) with regards to the quality of learned

transformation as well as the robustness under different training conditions.

Our novel contributions in the proposed work include:

- We propose an unsupervised learning framework which incorporates the Sinkhorn distance as a distributional similarity measure in the back-translation loss function.
- We use a neural network to optimize our model, especially to implement the Sinkhorn distance whose calculation itself is an optimization problem.
- Unlike previous models which only consider cross-lingual transformation in a single direction, our model jointly learns the word embedding transfer in both directions for each language pair.
- We present an intensive comparative evaluation where our model achieved the state-of-the-art performance for many language pairs in cross-lingual tasks.

## 2 Related Work

We divide the related work into supervised and unsupervised categories. Representative methods in both categories are included in our comparative evaluation (Section 3.4). We also discuss some related work in unsupervised domain transfer in addition.

**Supervised Methods:** There is a rich body of supervised methods for learning cross-lingual transfer of word embeddings based on bilingual dictionaries (Mikolov et al., 2013; Faruqui and Dyer, 2014; Artetxe et al., 2016; Xing et al., 2015; Duong et al., 2016; Gouws and Sogaard, 2015), sentence-aligned corpora (Kočiskỳ et al., 2014; Hermann and Blunsom, 2014; Gouws et al., 2015) and document-aligned corpora (Vulić and Moens, 2016; Sogaard et al., 2015). The most relevant line of work is that by Mikolov et al. (2013) where they showed monolingual word embeddings are likely to share similar geometric properties across languages although they are trained separately and hence cross-lingual mapping can be captured by a linear transformation across embedding spaces. Several follow-up studies tried to improve the cross-lingual transformation in various ways (Faruqui and Dyer, 2014; Artetxe et al., 2016; Xing et al., 2015; Duong et al., 2016; Ammar et al., 2016; Artetxe et al., 2016; Zhang et al.,

2016; Shigeto et al., 2015). Nevertheless, all these methods require bilingual lexicons for supervised learning. Vulić and Korhonen (2016) showed that 5000 high-quality bilingual lexicons are sufficient for learning a reasonable cross-lingual mapping.

**Unsupervised Methods** have been studied to establish cross-lingual mapping without any human-annotated supervision. Earlier work simply relied on word occurrence information only (Rapp, 1995; Fung, 1996) while later efforts have considered more sophisticated statistics in addition (Haghighi et al., 2008). The main difficulty in unsupervised learning of cross-lingual mapping is the formulation of the objective function, i.e., how to measure the goodness of an induced mapping without any supervision is a non-trivial question. Cao et al. (2016) tried to match the mean and standard deviation of the embedded word vectors in two different languages after mapping the words in the source language to the target language. However, such an approach has shown to be sub-optimal because the objective function only carries the first and second order statistics of the mapping. Artetxe et al. (2017) tried to impose an orthogonal constraint to their linear transformation model and minimize the distance between the transferred source-word embedding and its nearest neighbor in the target embedding space. Their method, however, requires a seed bilingual dictionary as the labeled training data and hence is not fully unsupervised. (Zhang et al., 2017a; Barone, 2016) adapted a generative adversarial network (GAN) to make the transferred embedding of each source-language word indistinguishable from its true translation in the target embedding space (Goodfellow et al., 2014). The adversarial model could be optimized in a purely unsupervised manner but is often suffered from unstable training, i.e. the adversarial learning does not always improve the performance over simpler baselines. Zhang et al. (2017b), Conneau et al. (2017) and Artetxe et al. (2017) also tried adversarial approaches for the induction of seed bilingual dictionaries, as a sub-problem in the cross-lingual transfer of word embedding.

**Unsupervised Domain Transfer:** Generally speaking, learning the cross-lingual transfer of word embedding can be viewed as a domain transfer problem, where the domains are word sets in different languages. Thus various work in the

field of *unsupervised domain adaptation* or *unsupervised transfer learning* can shed light on our problem. For example, He et al. (2016) proposed a semi-supervised method for machine translation to utilize large monolingual corpora. Shen et al. (2017) used unsupervised learning to transfer sentences of different sentiments. Recent work in computer vision addresses the problem of image style transfer without any annotated training data (Zhu et al., 2017; Taigman et al., 2016; Yi et al., 2017). Among those, our work is mostly inspired by the work on CycleGAN (Zhu et al., 2017), and we adopt their cycled consistent loss over images into our back-translation loss. One key difference of our method from CycleGAN is that they used the training loss of an adversarial classifier as an indicator of the distributional distance, but instead, we introduce the Sinkhorn distance in our objective function and demonstrate its superiority over the representative method using adversarial loss (Zhang et al., 2017a).

### 3 Proposed Method

Our system takes two sets of monolingual word embeddings of dimension  $d$  as input, which are trained separately on two languages. We denote them as  $X = \{x_i\}_{i=1}^n, Y = \{y_j\}_{j=1}^m, x_i, y_j \in \mathbb{R}^d$ . During the training of monolingual word embedding for  $X$  and  $Y$ , we also have the access to the word frequencies, represented by vectors  $r \in \mathbb{N}^n$  and  $c \in \mathbb{N}^m$  for  $X$  and  $Y$ , respectively. Specifically,  $r_i$  is the frequency for word (embedding)  $x_i$  and similarly for  $c_j$  of  $y_j$ . As illustrated in Figure 3, our model has two mappings:  $G : X \rightarrow Y$  and  $F : Y \rightarrow X$ . We further denote transferred embedding from  $X$  as  $G(X) := \{G(x_i)\}_{i=1}^n$  and correspondingly for  $F(Y)$ .

In the unsupervised setting, the goal is to learn the mapping  $G$  and  $F$  without any paired word translation. To achieve this, our loss function consists of two parts: Sinkhorn distance (Cuturi, 2013) for matching the distribution of transferred embedding to its target embedding distribution; and a back-translation loss for preventing degenerated transformation.

#### 3.1 Sinkhorn Distance

##### 3.1.1 Definition

Sinkhorn distance is a recently proposed distance between probability distributions. We use the Sinkhorn distance to measure the closeness be-

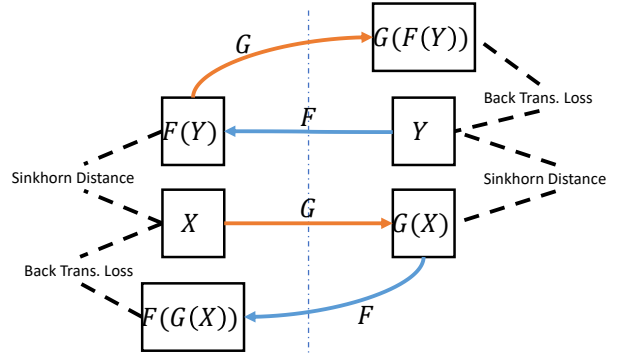


Figure 1: The model takes monolingual word embedding  $X$  and  $Y$  as input.  $G$  and  $F$  are embedding transfer functions parameterized by a neural network, which are represented by solid arrows. The dashed lines indicate the input for our objective losses, namely the Sinkhorn distance and back-translation loss

tween  $G(X)$  and  $Y$ , and also between  $F(Y)$  and  $X$ . During the training, our model optimizes  $G$  and  $F$  for lower Sinkhorn distance to make the transferred embeddings match the distribution of the target embeddings. Here we only illustrate the Sinkhorn distance between  $G(X)$  and  $Y$ , the derivation for  $F(Y)$  and  $X$  is very similar. Although the vocabulary sizes of two languages could be different, we are able to sample mini-batches of equal size from  $G(X)$  and  $Y$ . therefore we assume  $n = m$  in the following derivation.

To compute Sinkhorn distance, we firstly compute a distance matrix  $M^{(G)} \in \mathbb{R}^{n \times m}$  between  $G(X)$  and  $Y$  where  $M_{ij}^{(G)}$  is the distance measure between  $G(x_i)$  and  $y_j$ . The superscript on  $M^{(G)}$  indicates the distance that depends on a parameterized transformation  $G$ . For instance, if we choose Euclidean distance as a measure (see Section 3.1.3 for more discussions), we will have

$$M_{ij}^{(G)} = \|G(x_i) - y_j\|_2.$$

Given the distance matrix, the Sinkhorn distance between  $P_{G(X)}$  and  $P_Y$  is defined as:

$$d_{sh}(G) := \min_{P \in U_\alpha(r, c)} \langle P, M^{(G)} \rangle \quad (1)$$

where  $\langle \cdot, \cdot \rangle$  is the Forbenius dot-product and  $U_\alpha(r, c)$  is an entropy constrained transport poly-

---

**Algorithm 1** Computation of Sinkhorn Distance  $d_{sh}(G)$ 


---

```

1: procedure SINKHORN( $M^{(G)}, r, c, \lambda, I$ )
2:    $K^{(G)} := e^{-\lambda M^{(G)}}$ 
3:    $v = \mathbf{1}_m/m$   $\triangleright$  normalized one vector
4:    $i = 0$ 
5:   while  $i < I$  do  $\triangleright$  iterate for  $I$  times
6:      $u = r./K^{(G)}v$ 
7:      $v = c./K^{(G)T}u$ 
8:      $i = i + 1$ 
9:    $d_{sh}(G) = u^T((K^{(G)} \otimes M^{(G)})v)$ 
10:  return  $d_{sh}(G)$   $\triangleright$  The Sinkhorn distance

```

---

tope, defined as

$$U_\alpha(r, c) = \{P \in \mathbb{R}_+^{n \times m} \mid P\mathbf{1}_m = r, P^T\mathbf{1}_n = c, h(P) \leq h(r) + h(c) - \alpha\} \quad (2)$$

Note that  $P$  is non-negative and the first two constraints make its element-wise sum be 1. Therefore,  $P$  can be seen as a set of probability distributions. The same applies for  $r$  and  $c$  since they are frequencies.  $h$  is the entropy function defined on any probability distributions and  $\alpha$  is a hyperparameter to choose. For any probabilistic matrix  $P \in U_\alpha(r, c)$ , it can be viewed as the joint probability of  $(G(X), Y)$ . The first two constraints ensure that  $P$  has marginal distribution on  $G(X)$  as  $P_{G(X)}$  and on  $Y$  as  $P_Y$ . We can also view  $P_{ij}$  as the evidence for establishing a translation between word vector  $x_i$  and word vector  $y_j$ .

An intuitive interpretation of equation (1) is that we are trying to find the optimal transport probability  $P$  under the entropy constraint such that the total distance to transport from  $G(X)$  to  $Y$  is minimized.

### 3.1.2 Computing Sinkhorn Distance $d_{sh}(G)$

Cuturi (2013) showed that the optimal solution of formula (1) has the form  $P^* = \text{diag}(u)K\text{diag}(v)$ , where  $u$  and  $v$  are some non-negative vectors and  $K^{(G)} := e^{-\lambda M^{(G)}}$ ;  $\lambda$  is the Lagrange multiplier for the entropic constraint in 2 and each  $\alpha$  in Equation (1) has one corresponding  $\lambda$ . The Sinkhorn distance can be efficiently computed by a matrix scaling algorithm. We present the pseudo code in Algorithm 1. Note that the computation of  $d_{sh}(G)$  only requires matrix-vector multiplication. Therefore, we can compute and back propagate the gradient of  $d_{sh}(G)$  with regards to the parameters in  $G$  using standard deep

learning libraries. We show our implementation details in Section 3.4 and supplementary material.

### 3.1.3 Choice of the Distance Metric

In Section 3.1.1, we used the Euclidean distance of vector pairs to define  $M^{(G)}$  and Sinkhorn distance  $d_{sh}(G)$ . However, in our preliminary experiment, we found that Euclidean distance of unnormalized vectors gave poor performance. Therefore, following the common practice, we normalize all word embedding vectors to have a unit L2 norm in the construction of  $M^{(G)}$ .

As pointed out in Theorem 1 of Cuturi (2013),  $M^{(G)}$  must be a valid metric in order to make  $d_{sh}(G)$  a valid metric. For example, the commonly used cosine distance, which is defined as  $\text{CosDist}(a, b) = 1 - \cos(a, b)$ , is not a valid metric because it does not satisfy triangle inequality<sup>1</sup>. Thus, for constructing  $M^{(G)}$ , we propose the square root cosine distance ( $\text{SqrtCosDist}$ ) below:

$$\text{SqrtCosDist}(a, b) := \sqrt{2 - 2\cos(a, b)} \quad (3)$$

$$M_{ij}^{(G)} = \text{SqrtCosDist}(G(x_i), y_j) \quad (4)$$

**Theorem 1.**  $\text{SqrtCosDist}$  is a valid metric.

*Proof.*  $\forall a, b \in \mathbb{R}^d$ , let  $\hat{a} = \frac{a}{\|a\|}$ ,  $\hat{b} = \frac{b}{\|b\|}$ . We have  $\cos(a, b) = \langle \hat{a}, \hat{b} \rangle$  and  $\langle \hat{a}, \hat{a} \rangle = \langle \hat{b}, \hat{b} \rangle = 1$ . Then

$$\begin{aligned} \text{SqrtCosDist}(a, b) &= \sqrt{2 - 2\cos(a, b)} \\ &= \sqrt{\langle \hat{a}, \hat{a} \rangle + \langle \hat{b}, \hat{b} \rangle - 2\langle \hat{a}, \hat{b} \rangle} \\ &= \sqrt{\langle \hat{a} - \hat{b}, \hat{a} - \hat{b} \rangle} \\ &= \|\hat{a} - \hat{b}\| \end{aligned}$$

Obviously, the last term is the Euclidean distance between normalized input vectors  $\hat{a}$  and  $\hat{b}$ . Since Euclidean distance is a valid metric, it follows that  $\text{SqrtCosDist}$  satisfies all the axioms for a valid metric.  $\square$

## 3.2 Objective Function

Given enough capacity,  $G$  is capable to transfer  $X$  to  $Y$  for arbitrary word-to-word mappings. To ensure that, we learn a meaningful translation and also to regularize the search space of possible transformations, we enforce the word embedding after the forward and the backward transformation

<sup>1</sup>If we select  $a = [1, 0]$ ,  $b = [\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$ ,  $c = [0, 1]$  We have  $\text{CosDist}(a, c) \geq \text{CosDist}(a, b) + \text{CosDist}(b, c)$ , which violates the triangle inequality.



should not diverge much from its original direction. We simply choose the back-translation loss based on the cosine similarity:

$$d_{bt}(G, F) = \sum_i 1 - \cos(x_i, F(G(x_i))) + \sum_j 1 - \cos(y_j, G(F(y_j))) \quad (5)$$

where  $\cos$  is the cosine similarity.

Putting everything together, we minimize the following objective function.

$$L_{X,Y,r,c}(G, F) = d_{sh}(G) + d_{sh}(F) + \beta d_{bt}(G, F) \quad (6)$$

where hyper-parameter  $\beta$  controls the relative weight of the last term against the first two terms in the objective function. By definition, computation of  $d_{sh}(G)$  or  $d_{sh}(F)$  involves another minimization problem as shown in Equation (1). We solve it using the matrix scaling algorithm in Section 3.1.2, and treat  $d_{sh}(G)$  as a deterministic and differentiable function of parameters in  $G$ . The same holds for  $d_{sh}(F)$  and  $F$ .

### 3.3 Wasserstein GAN Training for Good Initial Point

In preliminary experiments, we found that our objective 6 is sensitive to the initialization of the weight in  $G$  and  $F$  in the purely unsupervised setting. It requires a good initial setting of the parameters to avoid getting stuck in the poor local minimal. To address this sensitivity issue, we employed a similar approach as in (Zhang et al., 2017b; Aldarmaki et al., 2018) to firstly used an adversarial training approach to learn  $G$  and  $F$  and use them as the initial point for training our full objective 6. More specifically, we choose to minimize the optimal transport distance below.

$$d_{ot}(G) := \min_{P \in U(r,c)} \langle P, M^{(G)} \rangle \quad (7)$$

$U$  is the transport polytope without entropy constraint, defined as follows.

$$U = \{P \in \mathbb{R}_+^{n \times m} | P \mathbf{1}_m = r, P^T \mathbf{1}_n = c\} \quad (8)$$

We optimize the distance above by its dual form and through adversarial training, which is also known as Wasserstein GAN (WGAN) (Arjovsky et al., 2017). We applied the optimization trick proposed by Gulrajani et al. (2017).

Although the first phase of adversarial training could be unstable, and the performance is lower than using the Sinkhorn distance, the adversarial training narrows down the search space of model parameters and boosting the training of our proposed model.

### 3.4 Implementation

We implemented transformation  $G$  and  $F$  by a linear transformation. The dimension of the input and output are the same with the word embedding dimension  $d$ .<sup>2</sup> For all the experiments in the subsequent section, the  $\beta$  in (6) was set to be 0.1. For hyper-parameters from the computation of Sinkhorn distance, we choose  $\lambda = 10$  and run the matrix scaling algorithm for 20 iterations. Due to the space constraint, a detailed implementation description is presented in the supplementary material. The code of our implementation is publicly available<sup>3</sup>.

## 4 Experiments

We conducted an evaluation of our approach in comparison with state-of-the-art supervised/unsupervised methods on several evaluation benchmarks for bilingual lexicon induction (Task 1) and word similarity prediction (Task 2). We include our main results in this section and report the ablation study in the supplementary material.

### 4.1 Data

#### 4.1.1 Monolingual Word Embedding Data

All the methods being evaluated in both tasks take monolingual word embedding in each language as the input data. We use publicly available pre-trained word embeddings trained on Wikipedia articles: (1) a smaller set of word embeddings of dimension 50 trained on comparable Wikipedia dump in five languages (Zhang et al., 2017a)<sup>4</sup> and (2) a larger set of word embeddings of dimension 300 trained on Wikipedia dump in 294 languages (Bojanowski et al., 2016)<sup>5</sup>. For conve-

<sup>2</sup>We tried more complex non-linear transformations for  $G$  and  $F$ . The performance is slightly worse than the linear case.

<sup>3</sup>Our implementation <https://github.com/xrc10/unsup-cross-lingual-embedding-transfer>

<sup>4</sup>Available at <http://nlp.csai.tsinghua.edu.cn/~zm/UBiLexAT>

<sup>5</sup>Available at <https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md>

nience, we name the two sets **WE-Z** and **WE-C**, respectively.

#### 4.1.2 Bilingual Lexicon Data

We need true translation pairs of words for evaluating methods in bilingual lexicon induction (Task 1). We followed previous studies and prepared two datasets below.

**LEX-Z:** Zhang et al. (2017a) constructed the bilingual lexicons from various resources. Since their ground truth word pairs are not released, we followed their procedure, crawled bilingual dictionaries and randomly separated them into the training and testing set of equal size.<sup>6</sup> Note that our proposed method did not utilize the training set. It was only used by supervised baseline methods described in Section 4.2. There are eight language pairs (order counted); the corresponding dataset statistics are summarized in Table 1. We use WE-Z embeddings in this dataset.

**LEX-C:** This lexicon was constructed by Conneau et al. (2017) and contains more translation pairs than LEX-Z. They divided them into training and testing set. We run our model and the baseline methods on 16 language pairs. For each language pair, the training set contains 5,000 unique query words and the testing set has 1,500 query words. We followed Conneau et al. (2017) and set the search space of candidate translations to be the 200,000 most frequent words in each target language. We use WE-C embeddings in this dataset.

#### 4.1.3 Bilingual Word Similarity Data

For bilingual word similarity prediction (Task 2) we need the true labels for evaluation. Following Conneau et al. (2017), we used the SemEval 2017 competition dataset, where human annotators measured the cross-lingual similarity of nominal word pairs according to the five-point Likert scale. This dataset contains word pairs across five languages: English (en), German (de), Spanish (es), Italian (it), and Farsi (fa). Each language pair has about 1,000 word pairs annotated with a real similarity score ranging from 0 to 4.

## 4.2 Baseline Methods

We evaluated the same set of supervised and unsupervised baselines for comparative evaluation in

<sup>6</sup>The bilingual dictionaries we crawled are submitted as supplementary material.

		# tokens	vocab. size	bi. lex. size
tr-en	tr	6m	7,482	18,404
	en	28m	13,220	27,327
es-en	es	61m	4,774	3,482
	en	95m	6,637	10,772
zh-en	zh	21m	3,349	54,170
	en	53m	5,154	51,375
it-en	it	73m	8,490	4,999
	en	93m	6,597	11,812

Table 1: The statistics of LEX-Z. The languages are Spanish (es), French (fr), Chinese (zh), Turkish (tr) and English (en). Number of tokens is the size of training corpus of WE-Z. The bilingual lexicon size means the number of unique words of a language in the gold bilingual lexicons.

both Task 1 and Task 2. The supervised baselines include the methods of Shigeto et al. (2015); Zhang et al. (2016); Artetxe et al. (2016); Xing et al. (2015); Mikolov et al. (2013); Artetxe et al. (2017).<sup>7</sup> We fed all the supervised methods with the bilingual dictionaries in the training portions of the LEX-Z and LEX-C datasets, respectively.

For unsupervised baselines we include the methods of Zhang et al. (2017a) and Conneau et al. (2017), whose source code is publicly available as provided by the authors.<sup>8</sup>

## 4.3 Results in Bilingual Lexicons Induction (Task 1)

Bilingual lexicon induction is a task to induce a translation in the target language for each query word in the source language. After the query word and the target-language words are represented in the same embedding space (or after our system maps the query word from the source embedding space to the target embedding space), the  $k$  nearest target words are retrieved based on their cosine similarity scores with respect to the query vector. If the  $k$  retrieved target words contain any valid translation according to the gold bilingual lexicon, the translation (retrieval) is considered successful. The fraction of the correctly translated source words in the test set is defined as  $accuracy@k$ ,

<sup>7</sup>The implementations are available from <https://github.com/artetxem/vecmap>.

<sup>8</sup>We used implementation by Zhang et al. (2017a) from <http://nlp.csai.tsinghua.edu.cn/~zm/UBiLexAT> and that of Conneau et al. (2017) from <https://github.com/facebookresearch/MUSE>

	Methods	tr-en	en-tr	es-en	en-es	zh-en	en-zh	it-en	en-it
Supervised	Mikolov et al. (2013)	19.41	10.81	68.73	41.19	45.88	<b>45.37</b>	59.83	<b>41.26</b>
	Zhang et al. (2016)	23.39	<b>11.07</b>	72.36	41.19	48.01	42.66	<b>63.19</b>	40.37
	Xing et al. (2015)	24.00	10.78	71.92	41.02	48.10	42.90	62.81	40.43
	Shigeto et al. (2015)	<b>26.56</b>	8.52	72.23	37.80	<b>49.95</b>	38.15	63.14	35.63
	Artetxe et al. (2016)	23.49	10.74	71.98	41.12	48.01	42.66	63.14	40.28
	Artetxe et al. (2017)	22.88	10.78	<b>72.61</b>	<b>41.62</b>	47.54	42.82	61.32	39.63
Unsupervised	Conneau et al. (2017)	4.09	1.41	60.16	33.58	41.98	34.70	26.98	15.47
	Zhang et al. (2017a)	15.83	7.41	63.41	37.73	42.08	41.26	54.75	37.17
	Ours	<b>23.29</b>	<b>9.96</b>	<b>73.05</b>	<b>41.95</b>	<b>49.03</b>	<b>44.63</b>	<b>61.42</b>	<b>39.63</b>

Table 2: The accuracy@k scores of all methods in bilingual lexicon induction on **LEX-Z**. The best score for each language pair is bold-faced for the supervised and unsupervised categories, respectively. Language pair "A-B" means query words are in language A and the search space of word translations is in language B. Languages are paired among **English(en)**, **Turkish (tr)**, **Spanish (es)**, **Chinese (zh)** and **Italian (it)**.

	Methods	bg-en	en-bg	ca-en	en-ca	sv-en	en-sv	lv-en	en-lv
Supervised	Mikolov et al. (2013)	44.80	<b>48.47</b>	57.73	<b>66.20</b>	43.73	<b>63.73</b>	26.53	<b>28.93</b>
	Zhang et al. (2016)	50.60	39.73	63.40	58.73	50.87	53.93	34.53	22.87
	Xing et al. (2015)	50.33	40.00	63.40	58.53	51.13	53.73	34.27	21.60
	Shigeto et al. (2015)	<b>61.00</b>	33.80	<b>69.33</b>	53.60	<b>61.27</b>	41.67	<b>42.20</b>	13.87
	Artetxe et al. (2016)	53.27	43.40	65.27	60.87	54.07	55.93	35.80	26.47
	Artetxe et al. (2017)	47.27	34.40	61.27	56.73	38.07	44.20	24.07	12.20
Unsupervised	Conneau et al. (2017)	26.47	13.87	41.00	33.07	24.27	24.47	-	-
	Zhang et al. (2017a)	-	-	-	-	-	-	-	-
	Ours	<b>50.33</b>	<b>34.27</b>	<b>58.60</b>	<b>54.60</b>	<b>48.13</b>	<b>50.47</b>	<b>27.73</b>	<b>13.53</b>

Table 3: The accuracy@k scores of all methods in bilingual lexicon induction on **LEX-C**. The best score for each language pair is bold-faced for the supervised and unsupervised categories, respectively. Languages are paired among **English(en)**, **Bulgarian(bg)**, **Catalan(ca)**, **Swedish(sv)** and **Latvian(lv)**. "-" means that during the training time, the model failed to converge to reasonable local minimal and hence the result is omitted in the table.

which is conventional metric in benchmark evaluations.

Table 2 shows the accuracy@1 for all the methods on LEX-Z in our evaluation. We can see that our method outperformed the other unsupervised baselines by a large margin on all the eight language pairs. Compared with the supervised methods, our method is still competitive (the best or the second-best scores on four out of eight language pairs), even ours does not require cross-lingual supervision. Also, we notice the performance variance over different language pairs. Our method outperforms all the methods (supervised and unsupervised combined) on the English-Spanish (en-es) pair, perhaps for the reasons that these two languages are most similar to each other, and that the monolingual word embeddings for this pair

in the comparable corpus are better aligned than the other language pairs. On the other hand, all the methods including ours have the worst performance on the English-Turkish (en-tr) pair. Another observation is the performance differences in the two directions of the language pair. For example, the performance of it-en is better than en-it for all methods in table 2. A part of the reason is that there are more unique English words than non-English words in the evaluation set. This would cause direction "xx-en" to be easier than "en-xx" because there are often multiple valid ground truth English translations for each query in "xx". But the same may not hold for the opposite direction of "en-xx". Nevertheless, the relative performance of our method compared to others is quite robust over different language pairs and different directions of

	Methods	de-en	en-de	es-en	en-es	fr-en	en-fr	it-en	en-it
Supervised	Mikolov et al. (2013)	61.93	<b>73.07</b>	74.00	<b>80.73</b>	71.33	<b>82.20</b>	68.93	<b>77.60</b>
	Zhang et al. (2016)	67.67	69.87	77.27	78.53	76.07	78.20	72.40	73.40
	Xing et al. (2015)	67.73	69.53	77.20	78.60	76.33	78.67	72.00	73.33
	Shigeto et al. (2015)	<b>71.07</b>	63.73	<b>81.07</b>	74.53	<b>79.93</b>	73.13	<b>76.47</b>	68.13
	Artetxe et al. (2016)	69.13	72.13	78.27	80.07	77.73	79.20	73.60	74.47
	Artetxe et al. (2017)	68.07	69.20	75.60	78.20	74.47	77.67	70.53	71.67
Unsupervised	Conneau et al. (2017)	<b>69.87</b>	<b>71.53</b>	<b>78.53</b>	79.40	<b>77.67</b>	<b>78.33</b>	<b>74.60</b>	<b>75.80</b>
	Zhang et al. (2017a)	-	-	-	-	-	-	-	-
	Ours	67.00	69.33	77.80	<b>79.53</b>	75.47	77.93	72.60	73.47

Table 4: The accuracy@k scores of all methods in bilingual lexicon induction on **LEX-C**. The best score for each language pair is bold-faced for the supervised and unsupervised categories, respectively. Languages are paired among **English (en)**, **German (de)**, **Spanish (es)**, **French (fr)** and **Italian (it)**. ”-” means that during the training time, the model failed to converge to reasonable local minimal and hence the result is omitted in the table.

translation.

Table 3 and Table 4 summarize the results of all the methods on the LEX-C dataset. Several points may be worth noticing. Firstly, the performance scores on LEX-C are not necessarily consistent with those on LEX-Z (Table 2) even if the methods and the language pairs are the same; this is not surprising as the two datasets differ in query words, word embedding quality, and training-set sizes. Secondly, the performance gap between the best supervised methods and the best unsupervised methods in both Table 3 and Table 4 are larger than that in Table 2. This is attributed to the large amount of good-quality supervision in LEX-C (5,000 human-annotated word pairs) and the larger candidate size in WE-C (200,000 candidates). Thirdly, the average performance in Table 3 is lower than that in Table 4, indicating that the language pairs in the former are more difficult than that in the latter. Nevertheless, we can see that our method has much stronger performance than other unsupervised methods in Table 3, i.e., on the harder language pairs, and that it performed comparably with the model by Conneau et al. (2017) in Table 4 on the easier language pairs. Combining all these observations, we see that our method is highly robust for various language pairs and under different training conditions.

#### 4.4 Results in Cross-lingual Word Similarity Prediction (Task 2)

We evaluate models on cross-lingual word similarity prediction (Task 2) to measure how much the predicted cross-language word similarities match

	Methods	de-en	es-en	fa-en	it-en
Supervised	Mikolov et al. (2013)	0.71	0.72	0.68	0.71
	Zhang et al. (2016)	0.71	0.71	0.69	0.71
	Xing et al. (2015)	0.72	0.71	0.69	0.72
	Shigeto et al. (2015)	0.72	<b>0.72</b>	0.69	0.71
	Artetxe et al. (2016)	<b>0.73</b>	<b>0.72</b>	<b>0.70</b>	<b>0.73</b>
	Artetxe et al. (2017)	0.70	0.70	0.67	0.71
Unsupervised	Conneau et al. (2017)	<b>0.71</b>	<b>0.71</b>	<b>0.68</b>	<b>0.71</b>
	Zhang et al. (2017a)	-	-	-	-
	Ours	<b>0.71</b>	<b>0.71</b>	0.67	<b>0.71</b>

Table 5: Performance (measured using Pearson correlation) of all the methods in cross-lingual semantic word similarity prediction on the benchmark data from Conneau et al. (2017). The best score in the supervised and unsupervised category is bold-faced, respectively. The languages include English (en), German (de), Spanish (es), Persian (fa) and Italian (it). ”-” means that the model failed to converge to reasonable local minimal during the training process.

the ground truth annotated by humans. Following the convention in benchmark evaluations for this task, we compute the Pearson correlation between the model-induced similarity scores and the human-annotated similarity scores over testing word pairs for each language pair. A higher correlation score with the ground truth represents the better quality of induced embeddings. All systems use the cosine similarity between the transformed embedding of each query and the word embedding of its paired translation as the predicted similarity score.

Table 5 summarizes the performance of all the methods in cross-lingual word similarity prediction. We can see that the unsupervised methods,



including ours, perform equally well as the supervised methods, which is highly encouraging.

## 5 Conclusion

In this paper, we presented a novel method for cross-lingual transformation of monolingual embeddings in an unsupervised manner. By simultaneously optimizing the bi-directional mappings w.r.t. Sinkhorn distances and back-translation losses on both ends, our model enjoys its prediction power as well as robustness, with the impressive performance on multiple evaluation benchmarks. For future work, we would like to extend this work in the semi-supervised setting where insufficient bilingual dictionaries are available.

## Acknowledgments

We thank the reviewers for their helpful comments. This work is supported in part by Defense Advanced Research Projects Agency Information Innovation Oce (I2O), the Low Resource Languages for Emergent Incidents (LORELEI) Program, Issued by DARPA/I2O under Contract No. HR0011-15-C-0114, and in part by the National Science Foundation (NSF) under grant IIS-1546329.

## References

- Aldarmaki, H., Mohan, M., and Diab, M. (2018). Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association for Computational Linguistics*, 6:185–196.
- Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.
- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.
- Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.
- Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462.
- Barone, A. V. M. (2016). Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv preprint arXiv:1608.02996*.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Cao, H., Zhao, T., Zhang, S., and Meng, Y. (2016). In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1818–1827.
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.
- Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*.
- Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.
- Fung, P. (1996). Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus A Non-parallel Corpus of Chinese and English. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.
- Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756.
- Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779.
- Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. *Proceedings of ACL-08: Hlt*, pages 771–779.

- He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W.-Y. (2016). Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.
- Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.
- Kočiský, T., Hermann, K. M., and Blunsom, P. (2014). Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*.
- Lu, A., Wang, W., Bansal, M., Gimpel, K., and Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322.
- Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6833–6844.
- Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., and Matsumoto, Y. (2015). Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151. Springer.
- Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.
- Søgaard, A., Agić, Ž., Alonso, H. M., Plank, B., Bohnet, B., and Johannsen, A. (2015). Inverted indexing for cross-lingual nlp. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*.
- Taigman, Y., Polyak, A., and Wolf, L. (2016). Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*.
- Vulić, I. and Korhonen, A. (2016). On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.
- Vulić, I. and Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.
- Xing, C., Wang, D., Liu, C., and Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.
- Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint*.
- Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017a). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1959–1970.
- Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017b). Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945.
- Zhang, Y., Gaddy, D., Barzilay, R., and Jaakkola, T. (2016). Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. Association for Computational Linguistics.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*.