Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction

Meng Zhang[†‡] **Yang Liu**[†‡] **Huanbo Luan**[†] **Maosong Sun**[†‡]
[†]State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[‡]Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China
zmlarry@foxmail.com, liuyang2011@tsinghua.edu.cn
luanhuanbo@gmail.com, sms@tsinghua.edu.cn

Abstract

Cross-lingual natural language processing hinges on the premise that there exists invariance across languages. At the word level, researchers have identified such invariance in the word embedding semantic spaces of different languages. However, in order to connect the separate spaces, cross-lingual supervision encoded in parallel data is typically required. In this paper, we attempt to establish the cross-lingual connection without relying on any cross-lingual supervision. By viewing word embedding spaces as distributions, we propose to minimize their earth mover's distance, a measure of divergence between distributions. We demonstrate the success on the unsupervised bilingual lexicon induction task. In addition, we reveal an interesting finding that the earth mover's distance shows potential as a measure of language difference.

1 Introduction

Despite tremendous variation and diversity, languages are believed to share something in common. Indeed, this belief forms the underlying basis of computational approaches to cross-lingual transfer (Täckström et al., 2013, *inter alia*), otherwise it would be inconceivable for the transfer to successfully generalize.

Linguistic universals manifest themselves at various levels of linguistic units. At the word level, there is evidence that different languages represent concepts with similar structure (Youn et al., 2016). Interestingly, as computational models of word semantics, monolingual word embeddings also exhibit isomorphism across languages (Mikolov et al., 2013a). This finding opens up the possibility to use a simple transformation, e.g. a linear map, to connect separately trained word embeddings cross-lingually. Learning such a transformation typically calls for cross-lingual supervision from parallel data (Faruqui and Dyer, 2014; Lu et al., 2015; Dinu et al., 2015; Lazaridou et al., 2015; Smith et al., 2017).

In this paper, we ask the question: Can we uncover the transformation without any cross-lingual supervision? At first sight, this task appears formidable, as it would imply that a bilingual semantic space can be constructed by using monolingual corpora only. On the other hand, the existence of structural isomorphism across monolingual embedding spaces points to the feasibility of this task: The transformation exists right there only to be discovered by the right tool.

We propose such a tool to answer the above question in the affirmative. The key insight is to view embedding spaces as distributions, and the desired transformation should make the two distributions close. This naturally calls for a measure of distribution closeness, for which we introduce the earth mover's distance. Therefore, our task can be formulated as the minimization of the earth mover's distance between the transformed source embedding distribution and the target one with respect to the transformation. Importantly, the minimization is performed at the distribution level, and hence no word-level supervision is required.

We demonstrate that the earth mover's distance minimization successfully uncovers the transformation for cross-lingual connection, as evidenced by experiments on the bilingual lexicon induction task. In fact, as an *unsupervised* approach, its performance turns out to be highly competitive with *supervised* methods. Moreover, as an interesting byproduct, the earth mover's distance provides a distance measure that may quantify a facet of language difference.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1934–1945
Copenhagen, Denmark, September 7–11, 2017.

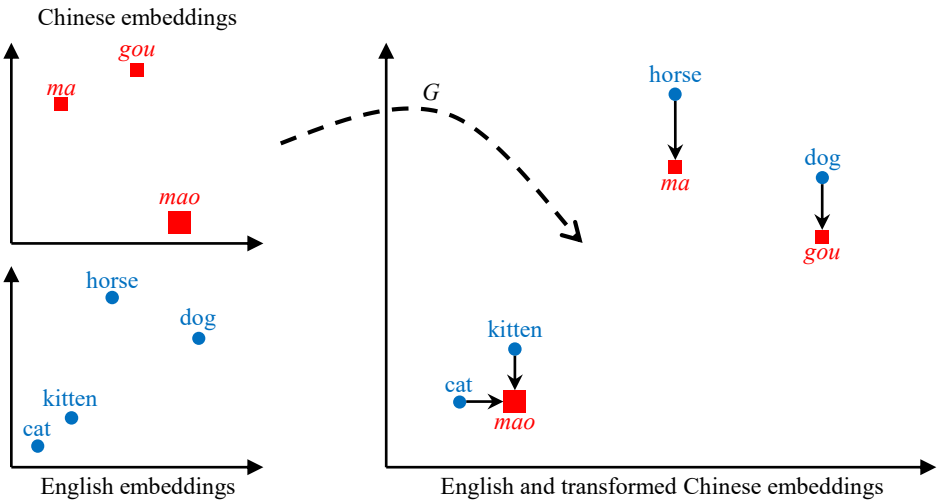

Figure 1: An illustration of our earth mover's distance minimization formulation. The subplots on the left schematically visualize Chinese and English embeddings. Due to isomorphism, there exists a simple transformation G that aligns the two embedding spaces well, as shown on the right. We expect to find the transformation G by minimizing the earth mover's distance without the need for cross-lingual word-level supervision, because the earth mover's distance holistically measures the closeness between two sets of weighted points. It computes the minimal cost of transporting one set of points to the other, whose weights are indicated by the sizes of squares and dots. We show the transport scheme in the right subplot with arrows, which can be interpreted as word translations.

2 Background

2.1 Aligning Isomorphic Embeddings

As discovered by previous work (Mikolov et al., 2013a), monolingual word embeddings exhibit isomorphism across languages, i.e., they appear similar in structure. However, as they are trained independently, the specific "orientation" of each embedding space is arbitrary, as illustrated in the left part of Figure 1. In order to connect the separate embedding spaces, we can try to transform the source embeddings so that they align well with target ones. Naturally, we need a measure for the quality of the alignment to guide our search for the transformation.

As we aim to eliminate the need for cross-lingual supervision from word translation pairs, the measure cannot be defined at the word level as in previous work (Mikolov et al., 2013a). Rather, it should quantify the difference between the entire distributions of embeddings. With this in mind, we find the earth mover's distance to be a suitable choice (Zhang et al., 2016b). Its workings are illustrated in the right part of Figure 1. We can think of target embeddings as piles of earth, and transformed source embeddings as holes to be filled. Then the earth mover's distance computes the minimal cost of moving the earth to fill the holes. Clearly, if the two sets of embeddings align well, the earth mover's distance will be small. Therefore, we can try to find the transformation that minimizes the earth mover's distance.

Another desirable feature of the earth mover's distance is that the computed transport scheme can be readily interpreted as translations. Moreover, this interpretation naturally handles multiple alternative translations. For example, the Chinese word "*mao*" can be translated to "cat" or "kitten", as shown in Figure 1.

2.2 The Form of the Transformation

The approximate isomorphism across embedding spaces inspires researchers to use a simple form of transformation. For example, Mikolov et al. (2013a) chose to use a linear transformation, i.e. the transformation G parametrized by a matrix. Later, proposals for using an orthogonal transformation are supported empirically (Xing et al., 2015; Zhang et al., 2016c; Artetxe et al., 2016) and theoretically (Smith et al., 2017). Indeed, an orthogonal transformation has desirable properties in this setting. If G is an orthogonal matrix that transforms the source embeddings into the target space, then its transpose (also its inverse) $G^\top$

performs transformation in the reverse direction. In that case, any word embedding a can be recovered by transforming back and forth because $G^\top Ga = a$. Moreover, computing the cosine similarity between a source embedding a and a target embedding b will be independent of the semantic space in which the similarity is measured, because $b^\top Ga / \|Ga\| \|b\| = a^\top G^\top b / \|a\| \left\|G^\top b\right\|$. Therefore we are inclined to use an orthogonal transformation for our task.

2.3 The Earth Mover's Distance

The earth mover's distance (EMD) is a powerful tool widely used in computer vision and natural language processing (Rubner et al., 1998; Kusner et al., 2015; Huang et al., 2016; Zhang et al., 2016b,a). Mathematically speaking, the EMD defines a distance between probability distributions. In the discrete case, a probability distribution can be represented by a sum of Dirac delta functions. For a pair of discrete distributions $\mathbb{P}_1 = \sum_i u_i \delta_{x_i}$ and $\mathbb{P}_2 = \sum_j v_j \delta_{y_j}$, the EMD is defined as

$$\mathrm{EMD}\left(\mathbb{P}_1, \mathbb{P}_2\right) = \min_{T \in \mathcal{U}(u,v)} \sum_i \sum_j T_{ij} c\left(x_i, y_j\right), \tag{1}$$

where $c\left(x_i, y_j\right)$ gives the ground distance between x_i and y_j, and $\mathcal{U}\left(u, v\right)$ is known as the transport polytope, defined as

$$\left\{ T | T_{ij} \geq 0, \sum_j T_{ij} = u_i, \sum_i T_{ij} = v_j, \forall i, j \right\}. \tag{2}$$

After solving the minimization program (1), the transport matrix T stores information of the transport scheme: A non-zero T_{ij} indicates the amount of probability mass transported from y_j to x_i. For our task, this can be interpreted as evidence for word translation (Zhang et al., 2016b), as indicated by arrows in the right part of Figure 1.

The EMD is closely related to the Wasserstein distance in mathematics, defined as

$$\mathrm{W}\left(\mathbb{P}_1, \mathbb{P}_2\right) = \inf_{\gamma \in \Gamma\left(\mathbb{P}_1, \mathbb{P}_2\right)} \mathbb{E}_{(x,y)\sim\gamma}\left[c\left(x, y\right)\right], \tag{3}$$

where $\Gamma\left(\mathbb{P}_1, \mathbb{P}_2\right)$ denotes the set of all joint distributions $\gamma\left(x, y\right)$ with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$ on the first and second factors respectively. As we can see, the Wasserstein distance generalizes the EMD to allow continuous distributions. In our context, we will use both terms interchangeably.

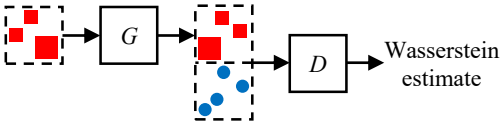

Figure 2: The Wasserstein GAN for unsupervised bilingual lexicon induction. The generator G transforms the source word embeddings into the target space. The critic D takes both sets of embeddings and tries to estimate their Wasserstein distance, and this information will be passed to the generator G during training to guide it towards minimizing the Wasserstein estimate.

3 Approaches

In our task, we are interested in a pair of distributions of word embeddings, one for the source language and the other for the target language. A source word embedding w_s^{S} is a d-dimensional column vector that represents the s-th source word in the V^{S}-sized source language vocabulary. Its distribution is characterized by a positive vector of frequencies f^{S} satisfying $\sum_{s=1}^{V^{\mathrm{S}}} f_s^{\mathrm{S}} = 1$, i.e. $P\left(w_s^{\mathrm{S}}\right) = f_s^{\mathrm{S}}$. Notations are similar for the target side. We assume the embeddings are normalized to have unit L_2 norm, which makes no difference to the result as we use cosine to measure semantic similarity.

Under this setting, we develop two approaches to our EMD minimization idea, called WGAN (Section 3.1) and EMDOT (Section 3.2) respectively.

3.1 Wasserstein GAN (WGAN)

Generative adversarial nets (GANs) are originally proposed to generate natural images (Goodfellow et al., 2014). They can generate sharp images if trained well, but they are notoriously difficult to train. Therefore, a lot of research efforts have been dedicated to the investigation into stabler training (Radford et al., 2015; Salimans et al., 2016; Nowozin et al., 2016; Metz et al., 2016; Poole et al., 2016; Arjovsky and Bottou, 2017), and the recently proposed Wasserstein GAN (Arjovsky et al., 2017) is a promising technique along this line of research.

While the original GAN is formulated as an adversarial game (hence its name), the Wasserstein GAN can be directly understood as minimizing the Wasserstein distance (3). Figure 2 illustrates the concept in the context of our unsupervised

bilingual lexicon induction task. The generator G takes source word embeddings and transforms them, with the goal that the transformed source distribution $\mathbb{P}^{G(\mathrm{S})}$ and the target distribution $\mathbb{P}^{\mathrm{T}}$ should be close as measured by the Wasserstein distance. The critic D takes both transformed source word embeddings and target word embeddings and attempts to accurately estimate their Wasserstein distance, which will guide the generator during training. The overall objective is

$$\min_{G \in \mathbb{R}^{d \times d}} \mathrm{W}\left(\mathbb{P}^{G(\mathrm{S})}, \mathbb{P}^{\mathrm{T}}\right), \tag{4}$$

where $\mathbb{P}^{G(\mathrm{S})} = \sum_{s=1}^{V^{\mathrm{S}}} f_s^{\mathrm{S}} \delta_{G w_s^{\mathrm{S}}}$ and $\mathbb{P}^{\mathrm{T}} = \sum_{t=1}^{V^{\mathrm{T}}} f_t^{\mathrm{T}} \delta_{w_t^{\mathrm{T}}}$ are the distributions of transformed source word embeddings and target word embeddings. Here we do not impose the orthogonal constraint on G to facilitate the use of a gradient-based optimizer. With the ground distance c being Euclidean distance L_2, the Kantorovich-Rubinstein duality (Villani, 2009) gives

$$\begin{aligned} &\mathrm{W}\left(\mathbb{P}^{G(\mathrm{S})}, \mathbb{P}^{\mathrm{T}}\right) \\ =&\frac{1}{K} \sup_{\|f\|_L \leq K} \mathbb{E}_{y \sim \mathbb{P}^{\mathrm{T}}}\left[f\left(y\right)\right] - \mathbb{E}_{y \sim \mathbb{P}^{G(\mathrm{S})}}\left[f\left(y\right)\right], \end{aligned} \tag{5}$$

where the supremum is over all K-Lipschitz functions f. As neural networks are universal function approximators (Hornik, 1991), we can attempt to approximate f with a neural network, called the critic D, with weight clipping to ensure the function family is K-Lipschitz. Therefore the objective of the critic is

$$\max_{D} \mathbb{E}_{y \sim \mathbb{P}^{\mathrm{T}}}\left[f_D\left(y\right)\right] - \mathbb{E}_{x \sim \mathbb{P}^{\mathrm{S}}}\left[f_D\left(Gx\right)\right]. \tag{6}$$

Conceptually, the critic D assigns scores f_D to real target embeddings and fake ones generated by the generator G. When the objective (6) is trained until optimality, the difference of the scores will approximate the Wasserstein distance up to a multiplicative constant. The generator G then aims to minimize the approximate distance, which leads to

$$\min_{G \in \mathbb{R}^{d \times d}} -\mathbb{E}_{x \sim \mathbb{P}^{\mathrm{S}}}\left[f_D\left(Gx\right)\right]. \tag{7}$$

3.2 EMD Minimization Under Orthogonal Transformation (EMDOT)

Alternative to minimizing the Wasserstein distance by duality, the primal program with the orthogonal constraint can be formalized as

$$\min_{G \in \mathcal{O}(d)} \mathrm{EMD}\left(\mathbb{P}^{G(\mathrm{S})}, \mathbb{P}^{\mathrm{T}}\right), \tag{8}$$

where $\mathcal{O}\left(d\right)$ is the orthogonal group in dimension d. The exact solution to this minimization program is NP-hard (Ding and Xu, 2016). Fortunately, an alternating minimization procedure is guaranteed to converge to a local minimum (Cohen and Guibas, 1999). Starting from an initial matrix $G^{(0)}$, we alternate between the following subprograms repeatedly:

$$T^{(k)} = \operatorname*{arg\,min}_{T \in \mathcal{U}(f^{\mathrm{S}}, f^{\mathrm{T}})} \sum_{s=1}^{V^{\mathrm{S}}} \sum_{t=1}^{V^{\mathrm{T}}} T_{st} c\left(G^{(k)} w_s^{\mathrm{S}}, w_t^{\mathrm{T}}\right), \tag{9}$$

$$G^{(k+1)} = \operatorname*{arg\,min}_{G \in \mathcal{O}(d)} \sum_{s=1}^{V^{\mathrm{S}}} \sum_{t=1}^{V^{\mathrm{T}}} T_{st}^{(k)} c\left(G w_s^{\mathrm{S}}, w_t^{\mathrm{T}}\right). \tag{10}$$

The minimization in (9) is the EMD program (1), with existing solvers available. For better scalability, we choose an approximate solver (Cuturi, 2013).

The minimization in (10) aims to find the transformation $G^{(k+1)}$ with cross-lingual connection provided in $T^{(k)}$. This is exactly the supervised scenario, and previous works typically resort to gradient-based solvers (Mikolov et al., 2013a). But they can be cumbersome especially as we impose the orthogonal constraint on G. Fortunately, if we choose the ground distance c to be the squared Euclidean distance L_2^2, the program (10) is an extension of the orthogonal Procrustes problem (Schönemann, 1966), which admits a closed-form solution:

$$G^{(k+1)} = UV^{\top}, \tag{11}$$

where U and V are obtained from a singular value decomposition (SVD):

$$\sum_{s=1}^{V^{\mathrm{S}}} \sum_{t=1}^{V^{\mathrm{T}}} T_{st}^{(k)} w_t^{\mathrm{T}} w_s^{\mathrm{S}\top} = USV^{\top}. \tag{12}$$

Note that the SVD is efficient because it is performed on a $d \times d$ matrix, which is typically low-dimensional. Choosing $c = L_2^2$ is also motivated by its equivalence to the cosine dissimilarity, as proved in Appendix A.

3.3 Discussion

Starting from the idea of earth mover's distance minimization, we have developed two approaches towards the goal. They employ different optimization techniques, which in turn lead to different

practical choices. For example, we choose $c = L_2^2$ for the EMDOT approach to obtain a closed-form solution to the subprogram (10), otherwise we would have to use gradient-based solvers. In contrast, the WGAN approach calls for $c = L_2$ because the Kantorovich-Rubinstein duality takes a simple form only in this case.

The EMDOT approach is attractive for several reasons: It is consistent for training and testing (the equivalence between the ground distance $c = L_2^2$ and cosine dissimilarity), compatible with the orthogonal constraint, mathematically sound (without much assumption and approximation), guaranteed to converge, almost hyperparameter free, and fast in speed (the alternating subprograms have either effective approximate solvers or closed-form solutions). However, it suffers from a serious limitation: The alternating minimization procedure only converges to local minima, and they often turn out to be rather poor in practice.

Although the WGAN approach employs a stochastic-gradient-based optimizer (RMSProp) and does not guarantee global optima either, it works reasonably well in practice. It seems better at exploring the parameter space and finally landing in a neighborhood of a good optimum. Like other success stories of using stochastic-gradient-based optimizers to train neural networks, theoretical understanding of the behavior remains elusive.

We can enjoy the best of both worlds by incorporating the merits of both approaches: First the WGAN approach locates a good neighborhood of the parameter space, and then, starting from a reasonable initialization, the EMDOT approach efficiently explores the neighborhood to achieve enhanced performance.

4 Experiments

We first investigate the learning behavior of our WGAN approach, and then present experiments on the bilingual lexicon induction task, followed by a showcase of the earth mover's distance as a language distance measure. Details of the data sets and hyperparameters are described in Appendices B and C.

4.1 Learning Behavior of WGAN

We analyze the learning behavior of WGAN by looking at a typical training trajectory on Chinese-English. During training, we save 100 models, translate based on the nearest neighbor, and record their accuracy as the bilingual lexicon induction performance indicator at these training checkpoints. In theory, the critic objective (6) provides an estimate of the Wasserstein distance up to a multiplicative constant, and a smaller Wasserstein distance should mean the transformed source embedding space and the target embedding space align better, which should in turn result in a better bilingual lexicon. This is validated in Figure 3 by the correlation between Wasserstein estimate and accuracy. Therefore, the Wasserstein estimate can serve as an indicator for the bilingual lexicon induction performance, and we can save the model with the lowest value during training as the final model.

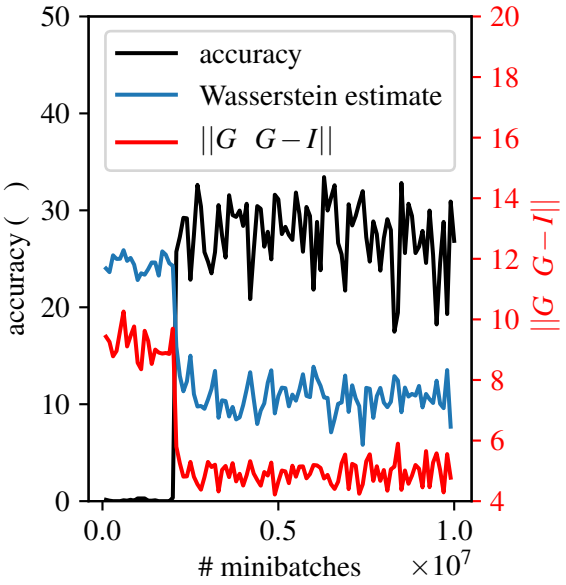

Figure 3: A typical training trajectory of WGAN. The three curves all correlate well. The Wasserstein estimate is rescaled because its magnitude is irrelevant.

In Figure 3, we also plot the value of $\left\|G^\top G - I\right\|_F$, which indicates the degree of orthogonality of the transformation matrix G. Interestingly, this also correlates nicely with the other curves, even though our WGAN formulation does not encourage G towards orthogonality. This finding confirms that a good transformation matrix is indeed close to orthogonality, and empirically justifies the orthogonal constraint for the EMDOT formulation.

Finally, we observe that the curves in Figure 3 are not very smooth. This means that although WGAN does well in exploring the parameter space and locating a reasonable transforma-

method	# seeds	zh-en	es-en	it-en	ja-zh	tr-en
TM	50	1.71	1.80	1.31	1.40	0.41
	100	17.27	24.93	23.22	22.91	15.02
	200	24.87	30.19	30.09	31.30	25.50
	500	28.24	32.11	31.69	35.79	32.63
IA	50	14.02	20.48	16.88	17.71	9.06
	100	22.14	28.73	25.99	28.24	18.37
	200	25.63	30.59	30.24	32.66	25.15
	500	27.21	31.94	31.54	35.33	31.50
WGAN	0	21.36	29.91	27.23	27.14	9.76
EMDOT	0	27.78	32.26	31.37	34.83	21.95

Table 1: F_1 scores for bilingual lexicon induction on Chinese-English, Spanish-English, Italian-English, Japanese-Chinese, and Turkish-English. The supervised methods TM and IA require seeds to train, and are listed for reference. Our EMDOT approach is initialized with the transformation found by WGAN, and consistently improves on it, reaching competitive performance with supervised methods.

tion matrix, it cannot stably refine the transformation. Fortunately, this is where EMDOT thrives, and hence combining them enjoys the benefits of both approaches.

4.2 Bilingual Lexicon Induction Performance

We test the quality of the cross-lingual transformation by evaluating on the bilingual lexicon induction task for five language pairs: Chinese-English, Spanish-English, Italian-English, Japanese-Chinese, and Turkish-English.

As the EMD automatically handles multiple alternative translations, we follow (Zhang et al., 2016b,a) to use F_1 score as the preferred evaluation metric.

Baselines

Our formulation is based on the isomorphism found across monolingual word embeddings. This idea has led to previous supervised methods:

- Translation matrix (TM) (Mikolov et al., 2013a): the pioneer of this type of methods, using linear transformation. We use a publicly available implementation.[1]

- Isometric alignment (IA) (Zhang et al., 2016c): an extension of TM by augmenting its learning objective with the isometric (orthogonal) constraint. Although Zhang et al. (2016c) had subsequent steps for their POS tagging task, it could be used for bilingual lexicon induction as well.

[1] http://clic.cimec.unitn.it/~georgiana.dinu/down

Although they need seed word translation pairs to train and thus not directly comparable to our system, we nonetheless report their results using {50, 100, 200, 500} seeds for a ballpark range of expected performance on this task, and skip the set of 500 seeds when testing all systems. We ensure the same input embeddings for these methods and ours. Their seeds are obtained through Google Translate (details in Appendix B.2). We apply the EMD as a postprocessing step (Zhang et al., 2016b) to allow them to handle multiple alternative translations. This is also done for our WGAN approach, as it does not produce the transport scheme to interpret as translation due to its duality formulation.

Results

Table 1 shows the F_1 scores on the five language pairs. As we can see, WGAN successfully finds a transformation that produces reasonable word translations. On top of that, EMDOT considerably improves the performance, which indicates that EMDOT refines the transformation found by WGAN.

Similar behavior across language pairs proves the generality of our approaches, as they build on embeddings learned from monolingual corpora without language-specific engineering. The quality of the embeddings, thus, will have an important effect on the performance, which may explain the lower scores on Turkish-English, as this low-resource setting may lack sufficient data to produce reliable embeddings. Higher noise levels in the preprocessing and ground truth for this lan-

	zh-en	es-en	it-en	ja-zh	tr-en
EMD	0.650	0.445	0.559	0.599	0.788
typology dissimilarity	0.467	0.342	0.259	0.433	0.541
geographical distance (km)	8161	1246	1464	2095	2854

Table 2: The earth mover's distance (EMD), typology dissimilarity, and geographical distance for Chinese-English, Spanish-English, Italian-English, Japanese-Chinese, and Turkish-English. The EMD shows correlation with both factors of linguistic difference.

guage pair (cf. the supplemental material), as well as the morphological richness of Turkish, may also be contributing factors to the relatively low scores.

Concerning the supervised methods `TM` and `IA`, they attain better performance with more supervision from seeds, as expected. For `TM` in particular, hundreds of seeds are needed for generalization, in line with the finding in (Vulić and Korhonen, 2016). Below that threshold, its performance drops dramatically, and this is when `IA` fares better with the orthogonal constraint. This indicates the importance of orthogonality when the seeds are few, or even zero as faced by our system. As the number of seeds increases, the performance of the supervised methods converges to a level comparable to our system.

4.3 The EMD as Language Distance

As our system minimizes the earth mover's distance between embeddings of two languages, we show here the final EMD can indicate the degree of difference between languages, serving as a proxy for language distance. Table 2 lists the EMD for the five language pairs considered in this paper, as well as their typology dissimilarity and geographical distance. The typology dissimilarity is computed from features in the WALS database (Dryer and Haspelmath, 2013). It is defined as one minus relative Hamming similarity, which is in turn defined as the number of agreeing features divided by the number of total features available for the language pair (Albu, 2006; Cysouw, 2013b). As a rough approximation, the geographical distance is measured by the distance between the capital cities of the countries where the considered languages are spoken (Eger et al., 2016).

The typology dissimilarity reflects genealogical influence on the divergence between languages, while the geographical distance indicates the effect of language contact. Both play important roles in shaping the languages we perceive today, and they also correlate with each other (Cysouw, 2013a). As we analyze Table 2, we find the EMD may be explained by both factors. Spanish-English and Italian-English are close both genealogically and geographically, and their EMD values are the lowest. English, Chinese, and Japanese belong to different language families, but the geographical proximity of the latter two enables intensive language contact, especially for the vocabularies, causing relatively smaller EMD. Finally, Turkish and English are distant in both aspects, and the EMD between them is large. Note that, however, the large EMD may also be caused by the relatively poor quality of monolingual embeddings due to low resource, and this should be a caveat of using the EMD to measure language distance.

5 Related Work

5.1 Bilingual Lexicon Induction

Bilingual lexicon induction is a long-standing research task in cross-lingual natural language processing. Traditional methods build statistical models for monolingual word co-occurrence, and combine cross-lingual supervision to solve the task. As word alignment for parallel sentences can produce fairly good bilingual lexica (Och and Ney, 2003), these methods focus on non-parallel data with a seed lexicon as cross-lingual supervision (Rapp, 1999; Gaussier et al., 2004).

An exception that does not rely on cross-lingual supervision is the decipherment approach (Dou and Knight, 2012, 2013; Dou et al., 2015). It views the source language as a cipher for the target language, and solves a statistical model that attempts to decipher the source language.

Following the popularity of monolingual word embeddings, cross-lingual word representation learning has also attracted significant attention in recent years. Building bilingual lexica from the learned cross-lingual embeddings is often considered an evaluative tool. Most methods rely on supervision encoded in parallel data, at the document

level (Vulić and Moens, 2015), the sentence level (Zou et al., 2013; Chandar A P et al., 2014; Hermann and Blunsom, 2014; Kočiský et al., 2014; Gouws et al., 2015; Luong et al., 2015; Coulmance et al., 2015; Oshikiri et al., 2016), or the word level (i.e. in the form of seed lexicon) (Gouws and Søgaard, 2015; Wick et al., 2016; Duong et al., 2016; Shi et al., 2015; Mikolov et al., 2013a; Faruqui and Dyer, 2014; Lu et al., 2015; Dinu et al., 2015; Lazaridou et al., 2015; Ammar et al., 2016; Zhang et al., 2016a, 2017; Smith et al., 2017).

There is a recent work that aims to remove the need for cross-lingual supervision (Cao et al., 2016). Similar to ours, the underlying idea is to match cross-lingually at the level of distribution rather than word. However, the distributions considered in that work are the hidden states of neural embedding models during the course of training. They are assumed to be Gaussian, so that the matching of distributions reduces to matching their means and variances, but this assumption is hard to justify and interpret. In contrast, our proposal does not make any assumption on the distributions, and directly matches the transformed source embedding distribution with the target distribution by minimizing their earth mover's distance.

Another attempt to learn cross-lingual embedding transformation without supervision is (Barone, 2016). Architectures of generative adversarial nets and adversarial autoencoders (Makhzani et al., 2015) are experimented, but the reported results are not positive. We tried the publicly available code on our data and obtained negative results as well. This outcome is likely caused by the training difficulty pointed out by (Arjovsky and Bottou, 2017), as traditional GAN training minimizes Jensen-Shannon divergence between distributions, which can provide pathological gradient to the generator and hamper its learning. The use of Wasserstein GAN addresses this problem and allows our simple architecture to be trained successfully.

5.2 Language Distance

Quantifying language difference is an open question with on-going efforts that put forward better measures based on manually compiled data (Albu, 2006; Hammarström and O'Connor, 2013). Researchers in computational linguistics also try to contribute corpus-based approaches to this question. Parallel data is typically exploited, and ideas range from information-theoretic (Juola, 1998), statistical (Mayer and Cysouw, 2012), to graph-based (Eger et al., 2016; Asgari and Mofrad, 2016). To our knowledge, the earth mover's distance is proposed for language distance for the first time, with the distinctive feature of relying on non-parallel data only.

5.3 The Earth Mover's Distance

First introduced into computer vision (Rubner et al., 1998), the earth mover's distance also finds application in natural language processing (Kusner et al., 2015; Huang et al., 2016), including bilingual lexicon induction (Zhang et al., 2016b,a). Zhang et al. (2016b) build upon bilingual word embeddings and apply the EMD program as a postprocessing step to automatically produce multiple alternative translations. Later, Zhang et al. (2016a) introduce the EMD into the training objective of bilingual word embeddings as a regularizer. These previous works rely on cross-lingual supervision, and do not approach the task from the view of embedding transformation, while our work formulates the task as EMD minimization to allow zero supervision.

Apart from the usage as a regularizer (Zhang et al., 2016a), the EMD can also play other roles in optimization programs designed for various applications (Cuturi and Doucet, 2014; Frogner et al., 2015; Montavon et al., 2016).

6 Conclusion and Future Work

In this work, we attack the problem of finding cross-lingual transformation between monolingual word embeddings in a purely unsupervised setting. We introduce earth mover's distance minimization to tackle this task by exploiting its distribution-level matching to sidestep the requirement for word-level cross-lingual supervision. Even though zero supervision poses a clear challenge, our system attains competitive performance with supervised methods for bilingual lexicon induction. In addition, the earth mover's distance provides a natural measure that may prove helpful for quantifying language difference.

We have implemented the earth mover's distance minimization framework from two paths, and their combination has worked well, but both can be potentially improved by recent advances

in optimization techniques (Gulrajani et al., 2017; Ding and Xu, 2016). Future work should also evaluate the earth mover's distance between more languages to assess its quality as language distance.

A Proof

The following proof shows that using squared Euclidean distance as the ground distance ($c = L_2^2$) is equivalent to using cosine dissimilarity when minimizing Equation (10).

$$\begin{aligned}
&\min_{G\in\mathcal{O}(d)} \sum_{s=1}^{V^{\mathrm{S}}}\sum_{t=1}^{V^{\mathrm{T}}} T_{st}^{(k)} \left\|Gw_s^{\mathrm{S}} - w_t^{\mathrm{T}}\right\|^2 \\
&= \min_{G\in\mathcal{O}(d)} \sum_{s=1}^{V^{\mathrm{S}}}\sum_{t=1}^{V^{\mathrm{T}}} T_{st}^{(k)} \\
&\quad \left(\left\|w_s^{\mathrm{S}}\right\|^2 + \left\|w_t^{\mathrm{T}}\right\|^2 - 2{w_t^{\mathrm{T}}}^{\top} G w_s^{\mathrm{S}}\right) \\
&= \min_{G\in\mathcal{O}(d)} -2\sum_{s=1}^{V^{\mathrm{S}}}\sum_{t=1}^{V^{\mathrm{T}}} T_{st}^{(k)} \cos\left(Gw_s^{\mathrm{S}}, w_t^{\mathrm{T}}\right) \\
&\quad + \text{const} \\
&= \min_{G\in\mathcal{O}(d)} -\sum_{s=1}^{V^{\mathrm{S}}}\sum_{t=1}^{V^{\mathrm{T}}} T_{st}^{(k)} \cos\left(Gw_s^{\mathrm{S}}, w_t^{\mathrm{T}}\right).
\end{aligned} \tag{13}$$

B Data Preparation

B.1 Non-Parallel Corpora for Training Embeddings

The data for training monolingual word embeddings comes from Wikipedia comparable corpora.[2] Following (Vulić and Moens, 2013), we retain only nouns with at least 1,000 occurrences except for Turkish-English, whose frequency cutoff threshold is 100, as the amount of data is relatively small in this low-resource setting. For the Chinese side, we first use OpenCC[3] to normalize characters to be simplified, and then perform Chinese word segmentation and POS tagging with THULAC.[4] The preprocessing of the English side involves tokenization, POS tagging, lemmatization, and lowercasing, which we carry out with the NLTK toolkit[5] for the Chinese-English pair. For Spanish-English and Italian-English, we choose to use TreeTagger[6] for preprocessing, as in (Vulić and Moens, 2013). For the Japanese corpus, we use MeCab[7] for word segmentation and POS tagging. For Turkish, we utilize the preprocessing tools (tokenization and POS tagging) provided in LORELEI Language Packs (Strassel and Tracey, 2016), and its English side is preprocessed by NLTK. The statistics of the preprocessed corpora is given in Table 3.

		# tokens	vocabulary size
zh-en	zh	21m	3,349
	en	53m	5,154
es-en	es	61m	4,774
	en	95m	6,637
it-en	it	73m	8,490
	en	93m	6,597
ja-zh	ja	38m	6,043
	zh	16m	2,814
tr-en	tr	6m	7,482
	en	28m	13,220

Table 3: Statistics of the non-parallel corpora for training monolingual word embeddings. Language codes: zh = Chinese, en = English, es = Spanish, it = Italian, ja = Japanese, tr = Turkish.

B.2 Seed Word Translation Pairs

The seed word translation pairs for the translation matrix (`TM`) and isometric alignment (`IA`) approaches are obtained as follows. First, we ask Google Translate[8] to translate the source language vocabulary. Then the target translations are queried again and translated back to the source language, and those that do not match the original source words are discarded. This helps to ensure the translation quality. Finally, the translations are discarded if they fall out of our target language vocabulary.

B.3 Ground Truth

As the ground truth bilingual lexicon for evaluation, we use Chinese-English Translation Lexicon Version 3.0 (LDC2002L27) for the Chinese-English pair. For Spanish-English and Italian-English, we access Open Multilingual WordNet[9] through NLTK. For Japanese-Chinese, we use an in-house lexicon. For Turkish-English, we build a set of ground truth translation pairs in the same

[2] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora
[3] https://github.com/BYVoid/OpenCC
[4] http://thulac.thunlp.org
[5] http://www.nltk.org
[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger
[7] http://taku910.github.io/mecab
[8] https://translate.google.com
[9] http://compling.hss.ntu.edu.sg/omw

way as how we obtain seed word translation pairs from Google Translate, described above.

C Hyperparameters

C.1 WGAN

We parametrize the critic D as a feed-forward neural network with one hidden layer of 500 neurons. The generator G is initialized with a random orthogonal matrix. The expectations in critic and generator objectives (6)(7) are approximated by minibatches of 1024 samples. We train for 10^7 minibatches. Most other hyperparameters follow from (Arjovsky et al., 2017) except the learning rates, for which larger values of 0.05 and 0.0005 are used for the generator and the critic respectively for faster convergence.

C.2 EMDOT

The approximate EMD solver (Cuturi, 2013) gives fairly accurate approximation with orders of magnitude speedup. However, it makes the transport matrix T no longer sparse. This is problematic, as we rely on interpreting a non-zero T_{st} as evidence to translate the s-th source word to the t-th target word (Zhang et al., 2016b). We therefore retain the largest pV^{S} elements of T, where p encodes our belief of the expected number of translations a source word can have. We set $p = 1.3$.

The alternating minimization procedure converges very fast. We run 10 iterations.

C.3 Monolingual Word Embeddings

As input monolingual word embeddings to the tested systems, we train the CBOW model (Mikolov et al., 2013b) with default hyperparameters in `word2vec`[10]. The embedding dimension d is 50.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work is supported by the National Natural Science Foundation of China (No. 61522204), the 973 Program (2014CB340501), and the National Natural Science Foundation of China (No. 61331013). This research is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme.

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

References

Mihai Albu. 2006. *Quantitative analyses of typological data*. Ph.D. thesis, Univ. Leipzig.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively Multilingual Word Embeddings. *arXiv:1602.01925 [cs]*.

Martin Arjovsky and Léon Bottou. 2017. Towards Principled Methods For Training Generative Adversarial Networks. In *ICLR*.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*.

Ehsaneddin Asgari and Mohammad R. K. Mofrad. 2016. Comparing Fifty Natural Languages and Twelve Genetic Languages Using Word Embedding Language Divergence (WELD) as a Quantitative Measure of Language Distance. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A Distribution-based Model to Learn Bilingual Word Embeddings. In *COLING*.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An Autoencoder Approach to Learning Bilingual Word Representations. In *NIPS*.

Scott Cohen and Leonidas Guibas. 1999. The Earth Mover's Distance Under Transformation Sets. In *ICCV*.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, Fast Cross-lingual Word-embeddings. In *EMNLP*.

Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *NIPS*.

Marco Cuturi and Arnaud Doucet. 2014. Fast Computation of Wasserstein Barycenters. In *ICML*.

Michael Cysouw. 2013a. Disentangling geography from genealogy. *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*.

Michael Cysouw. 2013b. Predicting language learning difficulty. *Approaches to measuring linguistic differences*.

Hu Ding and Jinhui Xu. 2016. FPTAS for Minimizing the Earth Mover's Distance Under Rigid Transformations and Related Problems. *Algorithmica*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. In *ICLR Workshop*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *EMNLP-CoNLL*.

Qing Dou and Kevin Knight. 2013. Dependency-Based Decipherment for Resource-Limited Machine Translation. In *EMNLP*.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying Bayesian Inference and Vector Space Models for Improved Decipherment. In *ACL-IJCNLP*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning Crosslingual Word Embeddings without Bilingual Corpora. In *EMNLP*.

Steffen Eger, Armin Hoenen, and Alexander Mehler. 2016. Language classification from bilingual word embedding graphs. In *COLING*.

Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *EACL*.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015. Learning with a Wasserstein Loss. In *NIPS*.

Eric Gaussier, J.M. Renders, I. Matveeva, C. Goutte, and H. Dejean. 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *ACL*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NIPS*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *ICML*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *NAACL-HLT*.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. *arXiv:1704.00028 [cs, stat]*.

Harald Hammarström and Loretta O'Connor. 2013. Dependency-sensitive typological distance. *Approaches to measuring linguistic differences*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Distributed Representations without Word Alignment. In *ICLR*.

Kurt Hornik. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks*.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised Word Mover's Distance. In *NIPS*.

Patrick Juola. 1998. Cross-Entropy and Linguistic Typology. In *CoNLL*.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning Bilingual Word Representations by Marginalizing Alignments. In *ACL*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In *ICML*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. In *ACL-IJCNLP*.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep Multilingual Correlation for Improved Word Embeddings. In *NAACL-HLT*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial Autoencoders. *arXiv:1511.05644 [cs]*.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. 2016. Unrolled Generative Adversarial Networks. *arXiv:1611.02163 [cs, stat]*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168 [cs]*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.

Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. 2016. Wasserstein Training of Restricted Boltzmann Machines. In *NIPS*.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. *arXiv:1606.00709 [cs, stat]*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *CL*.

Takamasa Oshikiri, Kazuki Fukui, and Hidetoshi Shimodaira. 2016. Cross-Lingual Word Representations via Spectral Graph Embeddings. In *ACL*.

Ben Poole, Alexander A. Alemi, Jascha Sohl-Dickstein, and Anelia Angelova. 2016. Improved generator objectives for GANs. *arXiv:1612.02780 [cs, stat]*. ArXiv: 1612.02780.

Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL*.

Y. Rubner, C. Tomasi, and L.J. Guibas. 1998. A Metric for Distributions with Applications to Image Databases. In *ICCV*.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *NIPS*.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning Cross-lingual Word Embeddings via Matrix Co-factorization. In *ACL-IJCNLP*.

Samuel Smith, David Turban, Steven Hamblin, and Nils Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR*.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In *LREC*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *TACL*.

Cédric Villani. 2009. *Optimal Transport: Old and New*.

Ivan Vulić and Anna Korhonen. 2016. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *ACL*.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses. In *NAACL-HLT*.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In *ACL-IJCNLP*.

Michael Wick, Pallika Kanani, and Adam Pocock. 2016. Minimally-Constrained Multilingual Embeddings via Artificial Code-Switching. In *AAAI*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *NAACL-HLT*.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*.

Meng Zhang, Yang Liu, Huanbo Luan, Yiqun Liu, and Maosong Sun. 2016a. Inducing Bilingual Lexica From Non-Parallel Data With Earth Mover's Distance Regularization. In *COLING*.

Meng Zhang, Yang Liu, Huanbo Luan, Maosong Sun, Tatsuya Izuha, and Jie Hao. 2016b. Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation. In *AAAI*.

Meng Zhang, Haoruo Peng, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Bilingual Lexicon Induction From Non-Parallel Data With Minimal Supervision. In *AAAI*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016c. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *NAACL-HLT*.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP*.