Point Set Registration for Unsupervised Bilingual Lexicon Induction

Hailong Cao and **Tiejun Zhao**
Harbin Institute of Technology
caohailong@hit.edu.cn, tjzhao@hit.edu.cn

Abstract

Inspired by the observation that word embeddings exhibit isomorphic structure across languages, we propose a novel method to induce a bilingual lexicon from only two sets of word embeddings, which are trained on monolingual source and target data respectively. This is achieved by formulating the task as point set registration which is a more general problem. We show that a transformation from the source to the target embedding space can be learned automatically without any form of cross-lingual supervision. By properly adapting a traditional point set registration model to make it be suitable for processing word embeddings, we achieved state-of-the-art performance on the unsupervised bilingual lexicon induction task. The point set registration problem has been well-studied and can be solved by many elegant models, we thus opened up a new opportunity to capture the universal lexical semantic structure across languages.

1 Introduction

Tremendous advances have been brought by distributed representations to the state-of-the-art natural language processing methods [Mikolov *et al.*, 2013b; Collobert and Weston, 2008; Pennington *et al.*, 2014]. In distributed representations, words are represented by real-valued points in a vector space referred to as word embeddings. Word embeddings learned automatically from monolingual data have the property that words with similar meaning are represented by points close to each other in the space.

It is natural to expect the above property still holds in the cross-lingual setting where words with similar meanings in different languages are represented by points close to each other in the shared embedding space. Learning word embeddings for cross-lingual natural language processing has attracted much attention [Klementiev *et al.*, 2012; Chandar *et al.*, 2014; Faruqui and Dyer, 2014; Hermann and Blunsom, 2014; Gouws *et al.*, 2015; Luong *et al.*, 2015; Shi *et al.*, 2015; Vulić and Moens, 2015; Upadhyay *et al.*, 2016; Smith *et al.*, 2017]. When used as the underlying input representation, cross-lingual word embeddings have been shown to boost the performance in many natural language processing tasks such as machine translation [Zou *et al.*, 2013; Zhang *et al.*, 2014] and transferring knowledge from high-resource languages to low-resource languages [Guo *et al.*, 2015], etc.

However, most of cross-lingual models require some form of cross-lingual supervision such as seed lexicon, word-level alignments, sentence-level alignments, document-level alignments and identical character strings shared by languages. Reliance on supervision might limit the development and application of cross-lingual representations. In this work, without requiring any form of cross-lingual supervision, we attempt to learn a transformation between two sets of word embeddings, which are trained on monolingual source and target data respectively. Our contributions include:

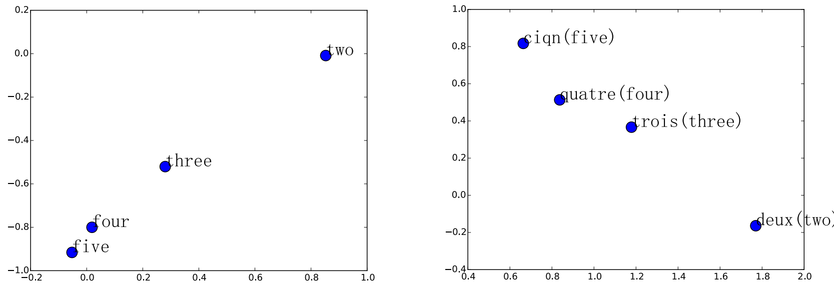
- We formulate our task as the point set registration problem [Myronenko and Song, 2010], which is a well studied more general problem. We thus open up a new unsupervised way for learning the correspondence between natural languages.
- Most point set registration models are designed for computer vision tasks which are quite different from the natural language processing problems. We show how to adapt a traditional point set registration algorithm to make it suitable for processing word embeddings.
- We achieved state-of-the-art performance on the task of unsupervised bilingual lexicon induction.

2 Unsupervised Bilingual Lexicon Induction

2.1 The problem

We start from two sets of word embeddings trained on monolingual source and target data respectively and assume their dimensionality is same. Throughout the paper, we use the following notations:

- D—dimensionality of monolingual word embeddings,
- N, M—vocabulary size of the target and source languages,
- $X_{N\times D} = (x_1, ..., x_N)^T$—the target word embeddings set,
- $Y_{M\times D} = (y_1, ..., y_M)^T$—the source word embeddings set,



(English) (French)

Figure 1: Distributed word vector representations of numbers in English (left) and French (right) learned separately from monolingual English and French data respectively. The four vectors in each language were projected down to two dimensions using PCA. If we left rotate the French vectors, it can be seen that these concepts have similar geometric arrangements in both spaces, suggesting that it is possible to unsupervised learn a transformation from one space to another. This is the key idea behind our method of unsupervised lexicon learning.

- $w(x), w(y)$—the frequency based weight of a target and a source word .

Given X, Y and the associated weights, our goal is to learn a transformation T (parameterized by a set of parameters θ) from the source to the target word embedding space so that for a source word embedding y, $T_\theta(y)$ lies close to the translation of y in the target embedding space. This is a very challenging task since there is no any form of cross-lingual supervision. Fortunately, [Mikolov *et al.*, 2013a] have found that natural languages have similar geometric arrangements in the embedding spaces. Figure 1 shows examples of English and French word embeddings. The similar geometric arrangements make it possible to learn a transformation in an unsupervised way. Intuitively, when one observed Figure 1, the most reasonable induced lexicon would contain these pairs: two-deux, three-trois, four-quatre and five-cinq. This is the reasoning behind our method.

2.2 Point Set Registration for Lexicon Induction

Taking X and Y as two sets of points in the D-dimensional space, we can formulate learning bilingual lexicon as the point set registration problem. This problem is well-studied in the field of computer vision and the goal of it is to assign correspondences between two sets of points and/or to recover the transformation that maps one point set to the other. Figure 2 shows an illustrative example of point set registration problem in the field of computer vision [Myronenko and Song, 2010].

If we compare the two examples in Figure 1 and Figure 2, we can find that the problem of learning bilingual lexicon based on distributed representation bears a striking resemblance to that of point set registration. In this work, we adopt a state-of-the-art point set registration method called Coherent Point Drift (CPD) algorithm proposed by [Myronenko and Song, 2010], which provides a suitable framework to our task at hand. The CPD algorithm considers the alignment of two point sets as a probability density estimation problem, where one point set represents the Gaussian mixture model (GMM) centroids and the other one represents the data points. We consider the points in transformed Y as the GMM centroids and the points in X as the data points generated by the GMM. The GMM probability density function is:

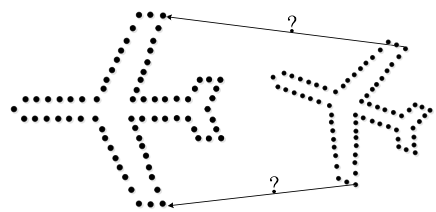


Figure 2: The point set registration problem: Given two sets of points, assign the correspondences and the transformation that maps one point set to the other [Myronenko and Song, 2010].

$$p(x) = \sum_{m=1}^{M+1} p(m)p(x|m) \tag{1}$$

where the first M component are:

$$p(x|m) = \frac{1}{(2\pi\sigma^2)^{D/2}} exp^{-\frac{\|x - T_\theta(y_m)\|^2}{2\sigma^2}} \tag{2}$$

where σ^2 is the equal isotropic covariance for all the M distributions. The $(M+1)^{th}$ component is an uniform distribution accounting for noise and outliers:

$$p(x|M+1) = \frac{1}{N} \tag{3}$$

The mixture weights are set as:

$$p(m) = \begin{cases} (1-\lambda)\frac{w(y_m)}{M} & 1 \le m \le M \\ \lambda & m = M+1 \end{cases} \tag{4}$$

where the weight λ reflects the priori assumption on the amount of noise in the point sets. In this work, we introduce the frequency based weights $w(y)$ and $w(x)$ to capture the intuition that the frequent source words are usually translated into frequent target words.

The parameter set θ of the transformation and the covariance σ^2 are estimated by minimizing the weighted negative log-likelihood of the N target words:

$$E(\theta, \sigma^2) = -\sum_{n=1}^{N} w(x_n) log \sum_{m=1}^{M+1} p(m)p(x_n|m) \quad (5)$$

The EM algorithm is used to find θ and σ^2. [Myronenko and Song, 2010] designed very elegant algorithms to efficiently implement the EM algorithm. Please refer to their paper for details.

3 Adaptation of the CPD Algorithm to Word Embeddings

Originally, most point set registration algorithms are designed for computer vision tasks. Mechanically applying them for natural language processing tasks might not achieve the optimal performance. Therefore, in this section, we describe how to adapt the CPD algorithm to make it suitable for processing word embeddings.

Specifically, [Myronenko and Song, 2010] proposed three kinds of point set registration methods which are based on rigid, affine and non-rigid transformations respectively. A rigid transformation only allows for translation, rotation, and scaling. Affine and non-rigid transformation allows anisotropic scaling and skews. In this paper, we choose to apply rigid point set registration which is based on orthogonal transformation. [Xing *et al.*, 2015; Artetxe *et al.*, 2016] have shown that the orthogonal transformation works well in word vector spaces. More recently, [Smith *et al.*, 2017] has proved that the optimal linear transformation between word vector spaces should be orthogonal. So in this paper we adopt the rigid point set registration and leave the affine and non-rigid transformation for future work.

3.1 Parameter Setting

For the rigid point set registration, the transformation is defined as:

$$T_\theta(y_m) = T_{R,t,s}(y_m) = sRy_m + t \quad (6)$$

where θ={R,t,s} and $R_{D\times D}$ is a orthogonal matrix, $t_{D\times 1}$ is a translation vector, and s is a scaling parameter.

Though [Myronenko and Song, 2010] has shown that such rigid transformation based method achieved very accurate result on their computer vision task, we have found that the translation vector t in Equation 6 is not so suitable for our problem which is defined on word embeddings(see Section 4.5). We conjecture there are two reasons for that. First, the translation vector t can freely move the data to any position in the high dimensional space and might make the model hard to train. Second, due to the introduction of t, the M step of the EM algorithm will normalize both X and Y by subtracting their means based on posterior probabilities. In this way, the position of each word relative to the origin will be changed and the angle between two word vectors will also be changed. Then the monolingual character of the word embeddings might not be held any more. So we suggest to remove the translation vector t from the original CPD algorithm and define the transformation as:

$$T_\theta(y_m) = T_{R,s}(y_m) = sRy_m \quad (7)$$

3.2 Initialization

The EM algorithm iteratively updates the parameters and the posterior probabilities. The posterior probability of the GMM centroid represented by the source word embedding y_m given the target word embedding x_n is defined as:

$$P(m|x_n) = \frac{exp^{-\frac{\|x_n - T_{R,s}(y_m)\|^2}{2\sigma^2}}}{\sum_{i=1}^{M} exp^{-\frac{\|x_n - T_{R,s}(y_i)\|^2}{2\sigma^2}} + c} \quad (8)$$

where $c = (2\pi\sigma^2)^{D/2}\frac{\lambda}{(1-\lambda)}\frac{M}{N}$. To calculate the posterior probability for the first time, we can guess the initial values of parameters based on prior knowledge. The EM algorithm is highly sensitive to initial values of model parameters. [Myronenko and Song, 2010] initialized the orthogonal matrix $R_{D\times D}$ as an identity matrix. This is assuming that there are one-to-one monotone correspondences between dimensions of the two spaces. It is reasonable for computer vision tasks since the rotation angle is usually not very large. However, our source and target word embeddings are trained entirely independently. So it is not reasonable to consider that first dimension in source space is more likely to correspond to the first dimension in target space and so on.

Instead of initializing the matrix $R_{D\times D}$, we propose to initialize posterior probabilities based on prior knowledge of languages. Intuitively, frequent source words are more likely to be translated into frequent target word, and vice-versa. Without loss of generality, we assume both word embeddings sets X and Y have already been sorted by the frequency of each word and define the initial posterior probability as:

$$P(m|x_n) = \frac{exp^{-\frac{(m-n\frac{M}{N})^2}{2\sigma_p^2}}}{\sum_{i=1}^{M} exp^{-\frac{(i-n\frac{M}{N})^2}{2\sigma_p^2}} + c} \quad (9)$$

where σ_p^2 is a prior variance. Based on this initial posterior probability we can estimate the initial values of parameters by the maximum step of the EM algorithm.

3.3 Data Normalization

A proper data normalization method can simplify the problem and achieve better performance. [Myronenko and Song, 2010] normalize both point sets to zero mean and unit variance before the registration. This works well for their computer vision task. But just like the translation vector $t_{D\times 1}$, such normalization method will also change the position of each word relative to the origin and the angle between two word vectors.

So instead of using the original normalization method in the CPD algorithm, we just normalize all embeddings to unit length. This is the most widely used normalization method for word embeddings in natural language processing tasks.

		# sentences	# words
en-fr	en	37,491,862	572,038,951
	fr	27,504,998	391,795,724
en-zh	en	15,592,216	313,854,226
	zh	4,453,302	134,490,433

Table 1: Statistics of the Wikipedia comparable corpora. Language codes: en = English, fr=French, zh = Chinese.

	top-1	top-5	top-10
MonoGiza without embeddings	0.19	0.38	NA
MonoGiza with embeddings	0.09	0.19	NA
[Zhang *et al.*, 2017]	51.91	65.10	69.88
Ours	53.34	67.30	71.03

Table 2: French-English top 1, top 5 and top 10 accuracies(%) of the MonoGiza, adversarial training and our method.

4 Experiments

In this section, we experimentally test the proposed model in comparison with related methods on the unsupervised bilingual lexicon induction task. We first train source and target word embeddings on source and target monolingual data independently using word2vec[1]. In detail, we used the CBOW model with negative sampling. The dimensionality of all word vectors is 50. The default values are used for all other parameters of word2vec. We retain only top 10k frequent words for each language. Second, we perform point set registration on the source and target word embeddings and obtain transformed source word embeddings. Third, we retrieve the top K nearest(measured by the cosine similarity) target word for each transformed source word as its translations and compare them against a ground truth lexicon. Following [Vulić and Moens, 2015], performance is measured by top K accuracy: If any of the K translations is found in the ground truth bilingual lexicon, the source word is considered to be correctly translated, and the accuracy is calculated as the percentage of correctly translated source words.

The C++ implementation of the CPD algorithm is available at https://github.com/gadomski/cpd. We adapt it for our task. For all hyper parameters in it, the default values are used. We set the parameter σ_p in Equation 9 as 100. There is not much differences when σ_p varies from 100 to 500. When we set the weight $w(y)$ and $w(x)$ in Equation 4 and 5, frequent words are penalized in a way similar to [Mikolov *et al.*, 2013b].

4.1 Data

We perform bilingual lexicon induction experiments on two language pairs. The first is French to English and the second is Chinese to English. The data for training monolingual word embeddings comes from Wikipedia comparable corpora [2]. The French and English text are tokenized and lowercased by scripts from www.statmt.org. All Chinese sentences are segmented by the Stanford Word Segmenter[3]. Table 1 lists the statistics of the final training data.

As the ground truth bilingual lexicons for evaluation, we use the lexicons derived by [Upadhyay *et al.*, 2016] using the Open Multilingual WordNet data released by [Bond and Foster, 2013] . The data includes synset alignments across 26 languages with over 90% accuracy. We pruned out words from each synset whose frequency rank is higher than 10k in the vocabulary of our training data and generated lexicons of sizes 1.1k and 1.4k pairs for en-fr and en-zh respectively.

4.2 Baselines

We compare our proposed method with two baselines both of which can induce bilingual lexicon from non-parallel data.

The first is a adversarial training [Goodfellow *et al.*, 2014] based method proposed by [Zhang *et al.*, 2017]. This is one of the best unsupervised lexicon induction model and moreover the code has been released[4]. Both their and our methods aim to learn a transformation between embedding spaces. But the transformations are learned in quite different ways.

The second baseline is a decipherment system based on a statistical model. The task of unsupervised bilingual lexicon induction has been referred as decipherment which has drawn significant amounts of interest in the past few years [Nuhn *et al.*, 2012; Dou *et al.*, 2015]. Decipherment views a foreign language as a cipher for English and finds a translation table that converts foreign texts into sensible English. In this work, we use MonoGiza[5] which implemented the state-of-the-art decipherment algorithms described in [Dou *et al.*, 2015] as the second baseline.

Both our model and that of [Zhang *et al.*, 2017] require pre-trained monolingual embeddings. MonoGiza can also utilize monolingual embeddings. To make sure that all results are comparable, we use the same embeddings as the input for all systems. Default values are used for all hyper parameters in both MonoGiza and the system of [Zhang *et al.*, 2017].

4.3 French to English Lexicon Induction

We first conduct French to English lexicon induction experiments. Table 2 lists the performance of the MonoGiza, the adversarial training based method and our point set registration based method. We report top 1, top 5 and top 10 accuracies. The powerful adversarial training is one of the best model for the task, the performance achieved by our method is comparable with that of adversarial training. Both adversarial training and point set registration based method leverage the isomorphic structure across languages and achieved better performance than the MonoGiza which is based on plain text.

The motivation examples shown in Figure 1 are just based on word embeddings used in this experiment. Here we show how the embeddings of the French numbers are transformed by the point set registration algorithm. Figure 3 shows these French and English numbers in the shared English embedding space in a 2D representation obtained by PCA. In the left part of the Figure 3, we simply merge all French and English embeddings trained by word2vec and then perform PCA on all

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

[2]http://linguatools.org/tools/corpora/wikipediacomparable-corpora

[3]http://nlp.stanford.edu/software/segmenter.shtml

[4]http://nlp.csai.tsinghua.edu.cn/ zm/UBiLexAT/

[5]http://www.isi.edu/natural-language/software/monogiza_release_v1.0.tar.gz

	top-1	top-5	top-10
MonoGiza without embeedings	0.14	0.28	NA
MonoGiza with embeddings	0.14	0.35	NA
[Zhang *et al.*, 2017]	38.82	57.29	64.27
Ours	39.75	60.02	66.49

Table 3: Chinese-English top 1, top 5 and top 10 accuracies(%) of the MonoGiza, adversarial training and our method.

embeddings. In the right part of the Figure 3, French embeddings are transformed by our method. It shows our method can capture the isomorphism exhibited by embeddings of languages and align them together.

Note that the two parts of Figure 1 are plotted in English and French embedding space respectively. While Figure 3 is plotted in the shared English embedding space where French and English embeddings are merged together. So Figure 1 and the left part of Figure 3 are a bit different.

4.4 Chinese to English Lexicon Induction

Now we describe Chinese to English lexicon induction experiments. Table 3 lists the performance of the MonoGiza, the adversarial training based method and our point set registration based method. The performance of our method is also very promising. Overall, the accuracies are lower than that for French-English. The reason might be that Chinese and English are relatively distantly related.

We visualized the Chinese and English embeddings in a 2D representation obtained by PCA and several examples are shown in Figure 4. The isomorphism of these words is not as obvious as that in Figure 3. Even though, when the Chinese embeddings are transformed by our method, some semantically equivalent Chinese and English words such as (beijing, 北京) or (stockholm, 斯德哥尔摩) are close to each other in the shared English embedding space.

4.5 The Importance of Adaptation Techniques

In section 3, we propose to adapt the CPD algorithm by removing the translation vector $t_{D\times 1}$, setting the initial value based on word frequency and normalizing the data to the unit hyper sphere. Table 4 lists the performance of applying the CPD algorithm with keeping all adaptation techniques and omitting one or all of them. The first row shows the performance of applying the CPD algorithm directly without any adaptation. The following three rows show the performances when the translation term is kept, original initialization and normalization methods are used respectively. Clearly, adaptation is quite necessary for the CPD algorithm to work for our task.

It is very surprising that the normalization methods make so much difference. Since the original "zero mean and unit variance"normalization will change the angle between two word embeddings, the cosine similarity might not be suitable for selecting Top K nearest translations. For a fair comparison, we also tried the Euclidean distance as the similarity measure for the original normalization case, without success.

	top-1	top-5	top-10
without adaptation	0.07	0.35	0.35
with translation vector	0.86	2.30	3.88
original initialization	0.00	0.28	0.28
original normalization	1.00	3.59	5.96
with adaptation	39.75	60.02	66.49

Table 4: Chinese-English top 1, top 5 and top 10 accuracies(%) of applying the CPD algorithm with keeping all adaptation techniques and omitting all or one of them.

4.6 Discussion

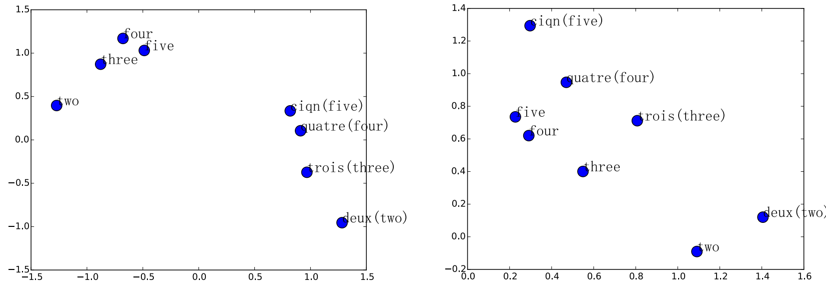
Essentially, both the method of [Zhang *et al.*, 2017] and ours are based on a linear transformation between embedding spaces. They learn the transformation by adversarial training which makes the target data and the transformed source data indistinguishable. Alternatively, our point set registration based method aims to maximizing the GMM probability for the target data conditioned on the transformed source data. Our experimental results show that the performance of the two methods are comparable. To further compare the two methods, we contrast the correct translation pairs induced by them respectively in the setting that only top 1 nearest target word is selected. In both French-English and Chinese-English cases, we found there are about 10 percent correct translation pairs are different from each other. This suggests that further improvements could be achieved if we can combine the advantages of them.

5 Related Work

This work is inspired by [Mikolov *et al.*, 2013a] who discovered that word embeddings learned separately from monolingual corpora exhibit similar geometrical structures across languages. Based on this interesting property, they proposed to learn a linear transformation from a source to a target embedding space. However, in their work, a seed bilingual lexicon is required to learn the transformation. In contrast, our method learns the transformation in an unsupervised way.

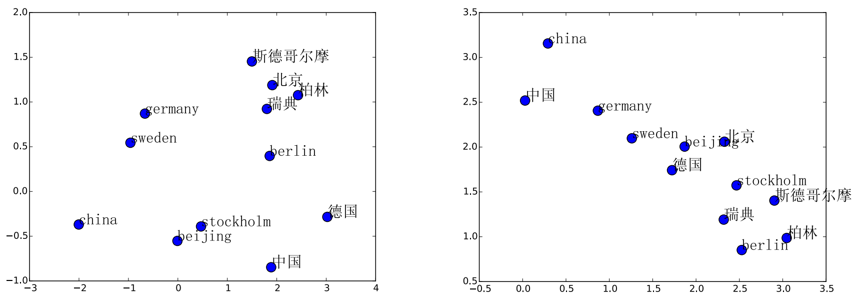
One of the most related work is the adversarial training based method proposed [Zhang *et al.*, 2017]. They designed very elegant loss functions and training techniques for adversarial training which is otherwise very hard to train. They made a breakthrough for unsupervised bilingual lexicon induction by a striking improvement over previous work. Essentially, both their and our methods are based on a linear transformation. They learn the transformation by adversarial training which makes the target data and the transformed source data indistinguishable. Alternatively, our method aims to maximizing the GMM probability for the target data conditioned on the transformed source data.

More recently, [Conneau *et al.*, 2017] further improved adversarial training based method by a refinement procedure. They first build a synthetic dictionary with adversarial training and then consider the most frequent words and retain only mutual nearest neighbors to ensure the quality. This high-quality dictionary is utilized to supervise their model to get much more improvements iteratively. In addition, they proposed a novel model selection method.



(Initial) (Transformed)

Figure 3: Word vector representations of English and French numbers in English embedding space. The four vectors in each language were projected down to two dimensions using PCA. The left part shows the initial results where both English and French embeddings are trained by word2vec, and in the right part the French embeddings are transformed by our method.



(Initial) (Transformed)

Figure 4: Word vector representations of English and Chinese words in English embedding space. The oracle translation pairs should be (china, 中国), (beijing, 北京), (germany, 德国), (berlin, 柏林), (sweden, 瑞典) and (stockholm, 斯德哥尔摩).

Both MonoGiza and our method are based on statistical models. MonoGiza also utilizes word embeddings and learns a transformation matrix between the two embedding spaces. Similarly, their transformation matrix is trained with stochastic EM. The main difference is that their model is based on plain text(word embeddings are used to make better generalization), while our method is based on vector spaces where the isomorphic structures across languages are leveraged.

[Xing *et al.*, 2015; Artetxe *et al.*, 2016] have shown that the orthogonal transformation works well in word vector spaces. More recently, [Smith *et al.*, 2017] has proved that the optimal linear transformation between word vector spaces should be orthogonal. An orthogonal transformation is also theoretically appealing for its self-consistency. However, in the case of unsupervised learning, imposing an orthogonal constraint to the transformation can make the model hard to optimize. [Zhang *et al.*, 2017] and [Conneau *et al.*, 2017] resorted to a loss function to encourage the transformation matrix to be close to an orthogonal matrix. In contrast, our method is based on a rigid point set registration algorithm which can obtain a strict orthogonal transformation and has a close-form solution.

[Artetxe *et al.*, 2017] proposed a self-learning framework which is able to learn high quality bilingual embeddings from as little bilingual evidence as a 25 word dictionary. We believe that their self-learning framework can be combined with our method to achieve more better performance.

6 Conclusions and Future Work

Inspired by the observation that word embeddings exhibit isomorphic structure across languages, we propose a novel method to induce a bilingual lexicon from only word embeddings trained on monolingual data. This is achieved by formulating the task as the point set registration problem. A transformation from the source to the target embedding space is learned automatically. By properly adapting a traditional point set registration model to make it suitable for our natural language task, we achieved very promising results for unsupervised bilingual lexicon induction. The point set registration problem has been well-studied in the field of computer vision and can be solved by many elegant models, we thus opened up a new opportunity to capture the universal lexical semantic structure across languages.

In the future work, we would further explore more advanced point set registration algorithms to obtain better performance for our task. In the opposite direction, it is also interesting to apply lexicon inducing methods such as adversarial training to point set registration.

Acknowledgments

We thank anonymous reviewers for their insightful comments. This work is funded by the projects of National Natural Science Foundation of China(No.61572154, No.71531013, No.91520204).

References

[Artetxe *et al.*, 2016] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*, pages 2289–2294, 2016.

[Artetxe *et al.*, 2017] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, pages 451–462, 2017.

[Bond and Foster, 2013] Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *ACL*, pages 1352–1362, 2013.

[Chandar *et al.*, 2014] A. P. Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In *NIPS*, pages 1853–1861, 2014.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167, 2008.

[Conneau *et al.*, 2017] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017.

[Dou *et al.*, 2015] Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. Unifying bayesian inference and vector space models for improved decipherment. In *ACL-IJCNLP*, pages 836–845, 2015.

[Faruqui and Dyer, 2014] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *EACL*, pages 462–471, 2014.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Gouws *et al.*, 2015] Stephan Gouws, Yoshua Bengio, and Greg Corado. Fast bilingual distributed representations without word alignments. In *ICML*, pages 748–756, 2015.

[Guo *et al.*, 2015] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. Cross-lingual dependency parsing based on distributed representations. In *ACL-IJCNLP*, pages 1234–1244, 2015.

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

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

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In *NAACL VSM*, pages 151–159, 2015.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv:1309.4168*, abs/1309.4168, 2013.

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

[Myronenko and Song, 2010] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drifts. *IEEE Transactions on PAMI*, 32(12):2262–2275, 2010.

[Nuhn *et al.*, 2012] Malte Nuhn, Arne Mauser, and Hermann Ney. Deciphering foreign language by combining language models and context vectors. In *ACL*, pages 156–164, 2012.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Shi *et al.*, 2015] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. Learning cross-lingual word embeddings via matrix co-factorization. In *ACL-IJCNLP*, pages 567–572, 2015.

[Smith *et al.*, 2017] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR*, 2017.

[Upadhyay *et al.*, 2016] Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. In *ACL*, pages 1661–1670, 2016.

[Vulić and Moens, 2015] Ivan Vulić and Marie-Francine Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL-IJCNLP*, pages 719–725, 2015.

[Xing *et al.*, 2015] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL-HLT*, pages 1006–1011, 2015.

[Zhang *et al.*, 2014] Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. Bilingually-constrained phrase embeddings for machine translation. In *ACL*, pages 111–121, 2014.

[Zhang *et al.*, 2017] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *ACL*, pages 1959–1970, 2017.

[Zou *et al.*, 2013] Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013.