Inducing Bilingual Lexica From Non-Parallel Data With Earth Mover's Distance Regularization

Meng Zhang[†‡] **Yang Liu**[†‡] **Huanbo Luan**[†] **Yiqun Liu**[†] **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, {yiqunliu, sms}@tsinghua.edu.cn

Abstract

Being able to induce word translations from non-parallel data is often a prerequisite for cross-lingual processing in resource-scarce languages and domains. Previous endeavors typically simplify this task by imposing the one-to-one translation assumption, which is too strong to hold for natural languages. We remove this constraint by introducing the Earth Mover's Distance into the training of bilingual word embeddings. In this way, we take advantage of its capability to handle multiple alternative word translations in a natural form of regularization. Our approach shows significant and consistent improvements across four language pairs. We also demonstrate that our approach is particularly preferable in resource-scarce settings as it only requires a minimal seed lexicon.

1 Introduction

Bilingual lexica provide word-level semantic equivalence information across languages, and prove to be valuable for a range of cross-lingual natural language processing tasks (Och and Ney, 2003; Levow et al., 2005; Täckström et al., 2013, *inter alia*). As building bilingual lexica from parallel corpora has been solved by word alignment (Och and Ney, 2003), researchers have turned their attention to non-parallel corpora. Accompanied by a small seed lexicon, non-parallel corpora are usually the only resources available in resource-scarce languages and domains, making the task of bilingual lexicon induction both important and challenging. A variety of statistical methods have been proposed to induce bilingual lexica from non-parallel data (Rapp, 1999; Koehn and Knight, 2002; Fung and Cheung, 2004; Gaussier et al., 2004; Haghighi et al., 2008; Ravi and Knight, 2011; Vulić et al., 2011; Vulić and Moens, 2013a; Vulić and Moens, 2013b; Dong et al., 2015). With the surge of word embeddings trained by neural networks, recent approaches that learn bilingual word representations from non-parallel data for bilingual lexicon induction have also shown promise (Mikolov et al., 2013b; Vulić and Moens, 2015).

However, none of the existing methods explicitly considers multiple alternative translation, i.e., the phenomenon that one source language word may have multiple possible translations in the target language. For example, the (romanized) Chinese word "*qiche*" can be translated to "car" or "automobile" in English, while the English word "car" can mean "*qiche*" or "*chexiang*" (railway carriage) in Chinese. Although prevalent among natural languages (Resnik and Yarowsky, 1999), multiple alternative translation is basically ignored by prior bilingual lexicon inducers; instead, they typically impose the one-to-one translation assumption (Vulić and Moens, 2013b) for simplicity. This represents a major drawback of existing bilingual lexicon induction approaches.

There has been one study that shows potential for tackling this issue. It introduces the Earth Mover's Distance (EMD) (Zhang et al., 2016). Given learned bilingual word embeddings, the EMD is used as a post-processing step to match vocabularies cross-lingually, which can be interpreted as word translation. Unlike the traditional K nearest neighbors leaving the determination of the number of translation proposals K to the user, the EMD automatically determines the list of translation candidates for each source word.

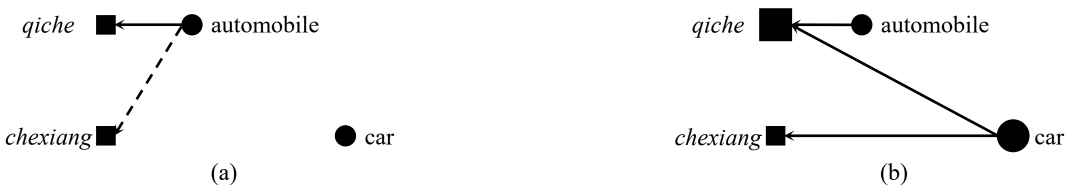

Figure 1: An illustration of bilingual word embeddings for translating from (romanized) Chinese to English. Arrows indicate translations, and solid ones are correct. (a) The nearest neighbor incorrectly translates "*chexiang*" to "automobile", and does not allow finding "car" as the other translation of "*qiche*". (b) The Earth Mover's Distance translates correctly. It associates words with weights, as indicated by the sizes of the shapes.

In this work, we propose to bring the EMD's capability to training. Intuitively, as the EMD in the post-processing step is able to connect a source word with multiple target word translations, it can play a more important role during training by driving the word vectors of these mutual translations to be closer. We therefore expect that the bilingual word embeddings learned this way will be more suitable for encoding multiple alternative translation by harnessing the power of the EMD. Our experiments validate the effectiveness of this strategy. A summary of our contributions is as follows:

- We introduce the Earth Mover's Distance into the training of bilingual word embeddings, and interpret it as a natural form of regularization for the overall learning objective (Section 3).
- We demonstrate significant and consistent performance improvement from our strategy across four language pairs (Sections 6.1 and 6.2).
- We investigate the effect of the number of seed word translation pairs, and find our approach to be most appealing with few seeds, in line with typical resource-scarce scenarios (Section 6.3).

2 Background

As an embedding-based approach to bilingual lexicon induction, the model consists of matrices $W^{\mathrm{S}} \in \mathbb{R}^{D \times V^{\mathrm{S}}}$ and $W^{\mathrm{T}} \in \mathbb{R}^{D \times V^{\mathrm{T}}}$, which pack up D-dimensional word embeddings of source and target languages with vocabulary sizes V^{S} and V^{T}, respectively. After training, these bilingual word embeddings are supposed to properly lie in the D-dimensional space that encodes cross-lingual semantic equivalence. To build a bilingual lexicon, or equivalently, to translate a source word into the target language, the nearest neighbor is typically employed to retrieve the target word embedding that is closest to the source word embedding.

The nearest neighbor has its limitations, as argued by Zhang et al. (2016). For one thing, the retrieval operation is inherently local (Figure 1(a)). Instead, they introduce the Earth Mover's Distance (EMD), which offers to match two sets of points with minimum total cost. For the word translation task, bilingual word embeddings can be naturally viewed as two sets of points. Therefore, given bilingual word embeddings, the EMD can perform word translation by providing optimal vocabulary-level matching in a holistic fashion (Figure 1(b)). This is achieved via the following optimization program:

$$\begin{aligned} &\min_{T} \sum_{t=1}^{V^{\mathrm{T}}} \sum_{s=1}^{V^{\mathrm{S}}} T_{ts} C_{ts} \\ &s.t.\ T_{ts} \geq 0 \\ &\quad \sum_{s=1}^{V^{\mathrm{S}}} T_{ts} \leq f_t^{\mathrm{T}}, t \in \{1, ..., V^{\mathrm{T}}\} \\ &\quad \sum_{t=1}^{V^{\mathrm{T}}} T_{ts} = f_s^{\mathrm{S}}, s \in \{1, ..., V^{\mathrm{S}}\} \end{aligned} \tag{1}$$

where C_{ts} defines the cost of matching the target word w_t^{T} and the source word w_s^{S} (illustrated by the distance between words in Figure 1), and f_t^{T} (resp. f_s^{S}) is the weight associated with w_t^{T} (resp. w_s^{S}) (illustrated by the sizes of the shapes in Figure 1(b)). The weights are chosen to be the number of times a word appears in the corpus. Once the linear program is solved, the matrix T stores the matching information between source and target vocabularies. This cross-lingual matching can be interpreted as translation. For example, a non-zero T_{ts} can be seen as evidence to translate the source word w_s^{S} to the target word w_t^{T}.

Besides the vocabulary-level matching, the EMD program brings an additional benefit. As mentioned in Section 1, it automatically retrieves multiple translations for a source word as long as the program finds it appropriate (cf. Figure 1). In the following section, we will strengthen this desirable capability by bringing the EMD program from a post-processing step to the training phase.

3 Approach

In typical scenarios, resources available to bilingual lexicon inducers include non-parallel corpora $\mathcal{C}^{\mathrm{S}}$ and $\mathcal{C}^{\mathrm{T}}$, and a seed lexicon $\mathbf{d}$. In order to utilize these resources to train bilingual word embeddings, a straightforward idea is to devise a learning objective that combines a monolingual term and a seed term.

The monolingual term $\mathcal{J}_{\mathrm{mono}}$ is responsible for explaining regularities in corpora $\mathcal{C}^{\mathrm{S}}$ and $\mathcal{C}^{\mathrm{T}}$. Since the two corpora are non-parallel, $\mathcal{J}_{\mathrm{mono}}$ consists of two monolingual submodels that are independent of each other:

$$\mathcal{J}_{\mathrm{mono}}\left(W^{\mathrm{S}}, W^{\mathrm{T}}\right) = \mathcal{J}_{\mathrm{mono}}^{\mathrm{S}}\left(W^{\mathrm{S}}\right) + \mathcal{J}_{\mathrm{mono}}^{\mathrm{T}}\left(W^{\mathrm{T}}\right). \tag{2}$$

As the common practice (Gouws et al., 2015), we choose the well established skip-gram model (Mikolov et al., 2013a) for our monolingual term.

The seed term $\mathcal{J}_{\mathrm{seed}}$ encourages embeddings of word translation pairs in a seed lexicon $\mathbf{d}$ to move near, which can be achieved via a L_2 regularizer:

$$\mathcal{J}_{\mathrm{seed}}\left(W^{\mathrm{S}}, W^{\mathrm{T}}\right) = -\sum_{\langle s,t\rangle \in \mathbf{d}} \left\| W_s^{\mathrm{S}} - W_t^{\mathrm{T}} \right\|^2, \tag{3}$$

where $s \in \left\{1, ..., V^{\mathrm{S}}\right\}$ and W_s^{S} is the s-th column of W^{S} (i.e. the embedding of the s-th source word w_s^{S}), and notations are similar for the target side.

However, as shown in our experiment, a simple linear combination of the monolingual term and the seed term is insufficient to provide satisfactory performance. We propose to introduce the Earth Mover's Distance into the training phase, as an additional term in the learning objective:

$$\mathcal{J}_{\mathrm{EMD}}\left(W^{\mathrm{S}}, W^{\mathrm{T}}, T\right) = -\sum_{t=1}^{V^{\mathrm{T}}} \sum_{s=1}^{V^{\mathrm{S}}} T_{ts} C_{ts} \tag{4}$$

with constraints

$$\begin{array}{c} T_{ts} \geq 0 \\ \sum_{s=1}^{V^{\mathrm{S}}} T_{ts} \leq f_t^{\mathrm{T}}, t \in \left\{1, ..., V^{\mathrm{T}}\right\} \\ \sum_{t=1}^{V^{\mathrm{T}}} T_{ts} = f_s^{\mathrm{S}}, s \in \left\{1, ..., V^{\mathrm{S}}\right\} \end{array}. \tag{5}$$

Note that, unlike the post-processing case (1), the ground distance matrix C is now parametrized by bilingual embeddings W^{S} and W^{T}, and therefore adjustable during training.

Putting everything together, we arrive at our overall learning objective to maximize:

$$\mathcal{J}\left(W^{\mathrm{S}}, W^{\mathrm{T}}, T\right) = \mathcal{J}_{\mathrm{mono}}\left(W^{\mathrm{S}}, W^{\mathrm{T}}\right) + \lambda_{\mathrm{s}} \mathcal{J}_{\mathrm{seed}}\left(W^{\mathrm{S}}, W^{\mathrm{T}}\right) + \lambda_{\mathrm{e}} \mathcal{J}_{\mathrm{EMD}}\left(W^{\mathrm{S}}, W^{\mathrm{T}}, T\right) \tag{6}$$

with constraints (5) inherited from the EMD. The hyperparameters λ_{s} and λ_{e} control the relative weighting of the terms. In this form, we can naturally view the EMD term as a regularizer that can potentially drive the embedding space to be more suitable for inducing bilingual lexica, especially multiple alternative word translation pairs.

The joint maximization of the overall learning objective (6) is clearly non-convex. In order to take advantage of the efficient solver specialized for the EMD program, we propose an alternating optimization procedure:

1. Fix W^{S} and W^{T}, and optimize with respect to T. This reduces to the usual linear EMD program with fixed ground distance, and the optimization can be achieved with the existing solver.

2. Fix T, and optimize with respect to W^{S} and W^{T}. Now the optimization can be easily achieved with stochastic gradient ascent.

4 Implementation

In this section, we describe details of a practical implementation of our approach.

4.1 Optimization

In our overall learning objective (6), unlike the other two, the EMD term $\mathcal{J}_{\mathrm{EMD}}$ requires an alternating optimization procedure. In order to allow each term to contribute to the learning process, we follow these steps. First, in each pass of the corpus (i.e. an epoch) the monolingual term $\mathcal{J}_{\mathrm{mono}}$ and the seed term $\mathcal{J}_{\mathrm{seed}}$ are optimized with asynchronous stochastic gradient ascent (Gouws et al., 2015). Then, we proceed to optimize the EMD term $\mathcal{J}_{\mathrm{EMD}}$ with the alternating optimization procedure. In Step 2 of the procedure, we take M gradient ascent steps. This hyperparameter is related to λ_{e}, as they jointly affect the strength of the EMD regularization. We are inclined to take small and many gradient ascent steps, so we fix $M = 10,000$ and tune λ_{e} on the validation set. Finally, the learning rate is decayed linearly at the end of each epoch.

4.2 Adding Context Vectors

In the previous section, we have presented our model with word vectors W^{S} and W^{T} as the parameters. In reality, each word is associated with a context vector as well (Mikolov et al., 2013c). While the usual representation of a word for evaluation is simply a word vector, some authors have suggested adding the context vector (Pennington et al., 2014; Levy et al., 2015). Previously this means a simple post-processing step during evaluation, but in our setting we can bring the trick to training. Specifically, using Euclidean distance as the ground distance, we would have parametrized C_{ts} in the EMD term (4) as

$$C_{ts} = \left\| W_t^{\mathrm{T}} - W_s^{\mathrm{S}} \right\| . \tag{7}$$

Considering the context vectors U^{S} and U^{T}, we now reformulate the ground distance as

$$C_{ts} = \left\| \left(W_t^{\mathrm{T}} + U_t^{\mathrm{T}} \right) - \left(W_s^{\mathrm{S}} + U_s^{\mathrm{S}} \right) \right\| . \tag{8}$$

This modification affects both steps in the alternating optimization procedure. In addition, the seed term also encourages corresponding context vectors to be close.

5 Experimental Setup

5.1 Data

In our experiments, the tested systems induce bilingual lexica from Wikipedia comparable corpora[1] on four language pairs: Chinese-English, Spanish-English, Italian-English, and Japanese-Chinese. Following (Vulić and Moens, 2013a), we retain only nouns that occur at least 1,000 times in our corpora.[2] For the Chinese side, we first use OpenCC[3] to normalize characters to be simplified, and then perform

[1] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora

[2] For Spanish-English and Italian-English, the cut-off frequency is 3,000 for a comparably-sized vocabulary.

[3] https://github.com/BYVoid/OpenCC

	zh-en		es-en		it-en		ja-zh	
	zh	en	es	en	it	en	ja	zh
# tokens	21M	53M	57M	90M	65M	88M	38M	16M
vocabulary size	3,349	5,154	2,543	3,557	3,378	3,534	6,043	2,814

Table 1: Training set statistics. Language codes: zh = Chinese, en = English, es = Spanish, it = Italian, ja = Japanese.

	zh-en	es-en	it-en	ja-zh
# test instances	1,938	1,860	2,051	2,320
# with multiple alternative translation	661	1,293	1,338	513

Table 2: Statistics of the test sets obtained by processing the gold standard lexica in the same way as (Zhang et al., 2016). A good portion of the test instances come with multiple alternative translation in the ground truth.

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, 2013a). For the Japanese corpus, we use MeCab[7] for word segmentation and POS tagging. The statistics of the preprocessed corpora is given in Table 1.

5.2 Seed Word Translation Pairs

We build our seed lexicon in a way similar to (Vulić and Moens, 2013a). First, we ask Google Translate[8] to translate the source side vocabulary. Then the translations in the target language are queried again in the reverse direction to translate back to the source language, and those that don't match with the original source words are discarded. This helps to ensure the quality of the translations. Finally, a translation pair is discarded if the target word falls out of our target vocabulary. We then take the most frequent S translation pairs as the seed lexicon. We vary S in our experiment to examine the effect of the seed lexicon size.

5.3 Evaluation Method

The limiting factor that prevents us from experimenting with truly resource-scarce language pairs is the unavailability of gold standard lexica for evaluation. Our focus on multiple alternative translation raises a higher demand that the gold standard lexica should include multiple possible translations for source words. For Chinese-English, we use Chinese-English Translation Lexicon Version 3.0[9] as the gold standard. For Spanish-English and Italian-English, we access Open Multilingual WordNet[10] through NLTK. For Japanese-Chinese, we use an in-house lexicon that meets our need. We reserve 10% of each gold standard lexicon for validation, and the remaining 90% for testing. We list test set statistics for each language pair in Table 2.

Following (Zhang et al., 2016), our evaluation metrics include accuracy A, precision P, recall R, and F_1 score. Accuracy is traditionally reported for the bilingual lexicon induction task, but it does not reflect the handling of multiple translations. This evaluative tradition also proves the lack of attention for multiple alternative translation. Therefore, we will be primarily looking at F_1 score in our experiments.

[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] https://catalog.ldc.upenn.edu/LDC2002L27

[10] http://compling.hss.ntu.edu.sg/omw

Method	A	P	R	F_1
STAT	0.2430	0.1589	0.1594	0.1591
MONO+SEED	0.2652	0.1983	0.1747	0.1858
Ours	0.5134	0.3770	0.3385	0.3567

Table 3: Performance on Chinese-English lexicon induction with 100 seed word translation pairs.

Method	Spanish-English	Italian-English	Japanese-Chinese
STAT	0.2384	0.2222	0.2117
MONO+SEED	0.2705	0.2350	0.1952
Ours	0.3686	0.3452	0.4111

Table 4: F_1 scores for three language pairs with 100 seed word translation pairs.

5.4 Baselines

We compare our approach to two baselines:

1. Statistics-based (STAT) (Gaussier et al., 2004).

2. Monolingual and seed terms (MONO+SEED).

The first baseline (STAT) is the traditional statistics-based approach, conventionally considered the standard approach to bilingual lexicon induction (Gaussier et al., 2004). It represents each word with a vector that encodes association strength between the word and seed words. We use a smoothed version of positive pointwise mutual information (PPMI) (Turney and Pantel, 2010) as the monolingual association measure.

The second baseline (MONO+SEED) is our system without the EMD term (i.e. $\lambda_{\mathrm{e}} = 0$). Comparison with it allows us to observe the effectiveness of the EMD term.

As we focus on multiple alternative translation but existing methods do not address it, we post-process the baselines by the EMD procedure (Zhang et al., 2016) to grant them the desired capability for a fair comparison with our approach.[11]

5.5 Hyperparameters

Our approach inherits hyperparameters from the monolingual skip-gram model, and includes term weights λ_{s} and λ_{e}. We set these hyperparameters based on tuning on the validation set, and observe little performance difference as long as they lie within a reasonable range. The monolingual hyperparameters are set as follows: embedding size D is 40; window size is 5; 5 negative samples; subsampling threshold is 10^{-5}; initial learning rate is 0.02; 20 training epochs. The statistics-based baseline uses a window size of 5 as well. The seed term weight λ_{s} is set to 0.01, and the EMD term weight λ_{e} is 0.0001.

6 Results

6.1 Performance on Chinese-English

We first report experimental results on Chinese-English lexicon induction with 100 seed word translation pairs, as shown in Table 3. We observe significant performance gains over both baselines, as measured by all evaluation metrics. In particular, comparing our approach with the MONO+SEED baseline highlights the effectiveness of introducing the EMD program into the training phase. As for training time, our approach takes about 4 hours, compared to 2 hours of MONO+SEED, due to the introduction of the EMD regularization.

[11] We found EMD post-processing to be generally superior to nearest neighbors, in line with (Zhang et al., 2016).

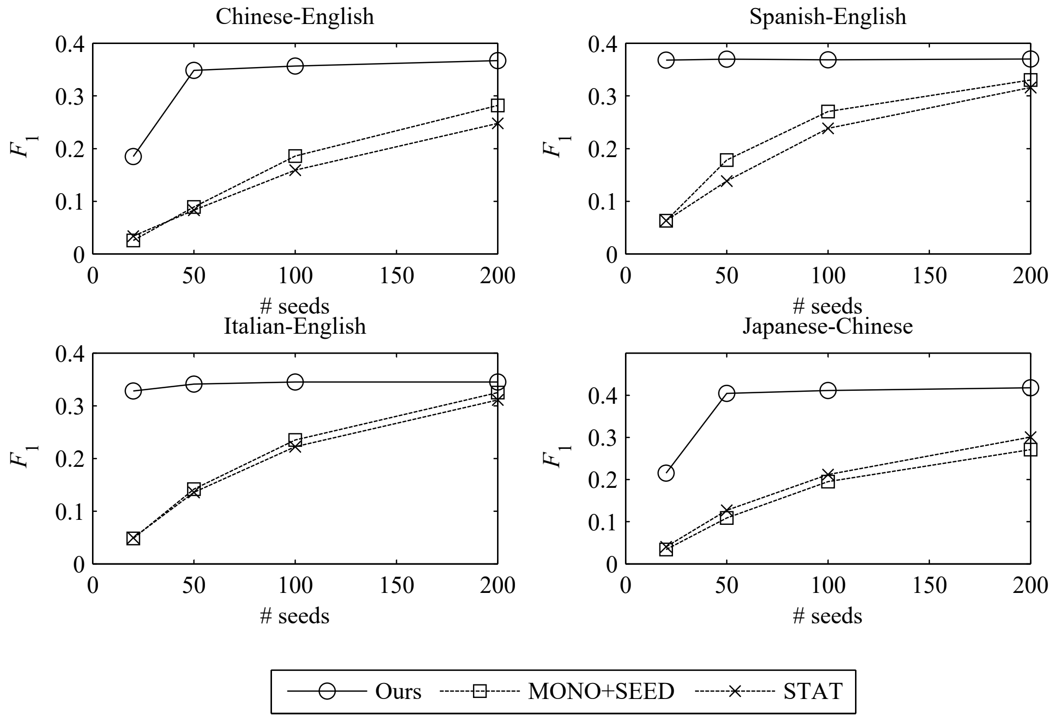

Figure 2: F_1 scores for the four language pairs with varying number of seed word translation pairs.

6.2 Performance on Other Language Pairs

We next experiment with the other three language pairs, i.e., Spanish-English, Italian-English, and Japanese-Chinese. The tested systems are provided with 100 seed word translation pairs as well. We only report the resulting F_1 scores in Table 4, as the other evaluation metrics exhibit similar trends. Relative to Chinese-English, these three language pairs should be more closely related. Nevertheless, the improvements of our method remain large, regardless of language pairs. The consistent performance signifies the generalizability of our approach across different language pairs.

6.3 Effect of Seed Lexicon Size

In this section, we investigate how the number of seed word translation pairs may affect the performance of the bilingual lexicon inducers. We vary the seed lexicon size in $\{20, 50, 100, 200\}$. Figure 2 shows the F_1 scores of the tested systems for the four language pairs. We observe that our system always attains high performance for the closely related language pairs Spanish-English and Italian-English, even when the seeds are as few as 20. For the more distant language pairs Chinese-English and Japanese-Chinese, 50 seeds suffice. In contrast, a limited number of seeds considerably degrades the performance of the baseline systems. Therefore, our system is particularly appealing in realistic resource-scarce scenarios for its minimal requirement for a seed lexicon, which is labor-intensive to compile.

6.4 Qualitative Analysis

In order to obtain a clearer view of the difference between the tested systems, we probe into the embeddings trained by them through a few examples. The Chinese-English translations in Table 5 imply that embeddings trained by our method appear superior. Although the two baselines may output more translations than our system, they often miss the correct ones, as shown by the examples of "*shan*" and "*jianzhu*". These translation differences should be eventually attributed to the quality of the underlying bilingual word embeddings, and in turn to the performance of the systems.

	STAT	MONO+SEED	Ours
qiche	good maker	**automobile** competitor customer luxury	**car** **automobile** **auto**
shan	palm flat waterfall dune barley chestnut citrus	middle part	**hill** **mountain**
jianzhu	architecture monument	foundation interior brick onwards	**building**

Table 5: English translations of three (romanized) Chinese words by the tested systems. The correct translations are in bold. The number of translations in each cell varies because it is automatically determined by the EMD program.

7 Related Work

Following its monolingual counterpart (Mikolov et al., 2013c, *inter alia*), bilingual word representation learning has attracted considerable attention. However, most of the works require parallel data as the cross-lingual signal (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), making them unsuitable for bilingual lexicon induction. Although a few exceptions exist (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Lu et al., 2015; Vulić and Moens, 2015; Shi et al., 2015; Gouws and Søgaard, 2015; Wick et al., 2016; Ammar et al., 2016), they lack a mechanism to deal with the multiple alternative translation prevalent cross-lingually.

The multiple alternative translation across languages is rooted in the polysemy of words within languages. In the monolingual setting, word sense disambiguation stands with a long line of research (Agirre and Rigau, 1996, *inter alia*). Since the advent of word representation learning, there have been some attempts to learn multiple vectors for a word, each dedicated to a single sense of the word, and therefore known as "sense embedding".

Existing sense embeddings can be roughly divided into two categories, depending on whether external resources are utilized. For those that do not rely on external resources, their main idea is to employ unsupervised methods like clustering to differentiate between multiple senses (Reisinger and Mooney, 2010; Huang et al., 2012; Neelakantan et al., 2014; Tian et al., 2014; Li and Jurafsky, 2015). For those that do, they typically retrofit existing word vectors to sense inventories (Jauhar et al., 2015; Rothe and Schütze, 2015), or use the resources to obtain a word sense disambiguation system, and then use it to disambiguate words, so that word representation learning methods can be applied (Chen et al., 2014; Iacobacci et al., 2015). An exception is the work of (Guo et al., 2014). Their external resource is parallel data. They observe that different senses of a word usually have different translations, so disambiguation can be thus achieved.

However, no prior research has shown how to connect sense embeddings cross-lingually, unless multilingual lexical ontologies exist (Camacho-Collados et al., 2015). For bilingual lexicon induction, where only non-parallel data and a seed lexicon are available, it is unclear whether sense embeddings can address multiple alternative translation.

Our work complements (Zhang et al., 2016): Their work applies the Earth Mover's Distance to the post-processing of fixed bilingual word embeddings to retrieve word translation, while ours strives to train better bilingual word embeddings with the EMD. In addition, we also explore the feasibility of using the EMD for bilingual lexicon induction from non-parallel data. In computer vision, there have been a few works that experiment with trainable ground distance in the EMD program (Wang and Guibas, 2012; Zen et al., 2014). However, they require supervision to properly guide the training. With supervision, their EMD program can stand alone to fit training data, while in our approach the EMD shows up as a regularizer in the learning objective. Besides, their models fully parametrize the ground distance as optimizable variables, whereas our model treats it as the Euclidean distance with adjustable word vectors.

8 Conclusion

In this paper, we look into multiple alternative translations prevalent across natural languages, which are largely neglected in previous bilingual lexicon induction research. We propose to introduce the Earth Mover's Distance into the training of bilingual word embeddings as a natural form of regularization. We provide strong empirical results for four language pairs to demonstrate the effectiveness of our approach. Furthermore, we discover that our method remains reliable with rather few seed word translation pairs, unlike the baselines exhibiting performance degradation. This advantage of our approach is particularly desirable in realistic resource-scarce settings.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work is supported by the 863 Program (2015AA015407), the National Natural Science Foundation of China (No. 61522204, No. 61432013 and No. 61303075), and Toshiba Corporation Corporate Research & Development Center.

References

Eneko Agirre and German Rigau. 1996. Word Sense Disambiguation Using Conceptual Density. In *COLING*.

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

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A Unified Multilingual Semantic Representation of Concepts. In *ACL-IJCNLP*.

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*.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A Unified Model for Word Sense Representation and Disambiguation. In *EMNLP*.

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

Meiping Dong, Yang Liu, Huanbo Luan, Maosong Sun, Tatsuya Izuha, and Dakun Zhang. 2015. Iterative Learning of Parallel Lexicons and Phrases from Non-Parallel Corpora. In *IJCAI*.

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

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *EMNLP*.

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*.

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

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

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources. In *COLING*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *ACL-HLT*.

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

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *ACL*.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *ACL-IJCNLP*.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models. In *NAACL-HLT*.

Philipp Koehn and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

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

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *TACL*.

Jiwei Li and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *EMNLP*.

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*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*.

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

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

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *EMNLP*.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

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

Sujith Ravi and Kevin Knight. 2011. Deciphering Foreign Language. In *ACL-HLT*.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-Prototype Vector-Space Models of Word Meaning. In *NAACL-HLT*.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering*.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *ACL-IJCNLP*.

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

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

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A Probabilistic Model for Learning Multi-Prototype Word Embeddings. In *COLING*.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *JAIR*.

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

Ivan Vulić and Marie-Francine Moens. 2013b. A Study on Bootstrapping Bilingual Vector Spaces from Non-Parallel Data (and Nothing Else). In *EMNLP*.

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

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying Word Translations from Comparable Corpora Using Latent Topic Models. In *ACL-HLT*.

Fan Wang and Leonidas J. Guibas. 2012. Supervised Earth Mover's Distance Learning and Its Computer Vision Applications. In *ECCV*.

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

Gloria Zen, Elisa Ricci, and Nicu Sebe. 2014. Simultaneous Ground Metric Learning and Matrix Factorization with Earth Mover's Distance. In *International Conference on Pattern Recognition*.

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

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