Adversarial Training 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

Word embeddings are well known to capture linguistic regularities of the language on which they are trained. Researchers also observe that these regularities can transfer across languages. However, previous endeavors to connect separate monolingual word embeddings typically require cross-lingual signals as supervision, either in the form of parallel corpus or seed lexicon. In this work, we show that such cross-lingual connection can actually be established without any form of supervision. We achieve this end by formulating the problem as a natural adversarial game, and investigating techniques that are crucial to successful training. We carry out evaluation on the unsupervised bilingual lexicon induction task. Even though this task appears intrinsically cross-lingual, we are able to demonstrate encouraging performance without any cross-lingual clues.

1 Introduction

As word is the basic unit of a language, the betterment of its representation has significant impact on various natural language processing tasks. Continuous word representations, commonly known as word embeddings, have formed the basis for numerous neural network models since their advent. Their popularity results from the performance boost they bring, which should in turn be attributed to the linguistic regularities they capture (Mikolov et al., 2013b).

Soon following the success on monolingual tasks, the potential of word embeddings for cross-lingual natural language processing has attracted much attention. In their pioneering work, Mikolov et al. (2013a) observe that word embeddings trained separately on monolingual corpora exhibit isomorphic structure across languages, as illustrated in Figure 1. This interesting finding is in line with research on human cognition (Youn et al., 2016). It also means a linear transformation may be established to connect word embedding spaces, allowing word feature transfer. This has far-reaching implication on low-resource scenarios (Daumé III and Jagarlamudi, 2011; Irvine and Callison-Burch, 2013), because word embeddings only require plain text to train, which is the most abundant form of linguistic resource.

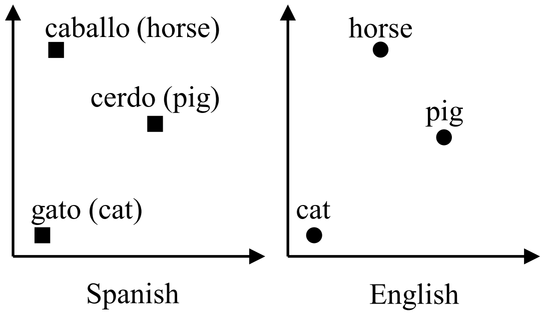

Figure 1: Illustrative monolingual word embeddings of Spanish and English, adapted from (Mikolov et al., 2013a). Although trained independently, the two sets of embeddings exhibit approximate isomorphism.

However, connecting separate word embedding spaces typically requires supervision from cross-lingual signals. For example, Mikolov et al. (2013a) use five thousand seed word translation pairs to train the linear transformation. In a recent study, Vulić and Korhonen (2016) show that at least hundreds of seed word translation pairs are needed for the model to generalize. This is unfortunate for low-resource languages and domains,

[*]Corresponding author.

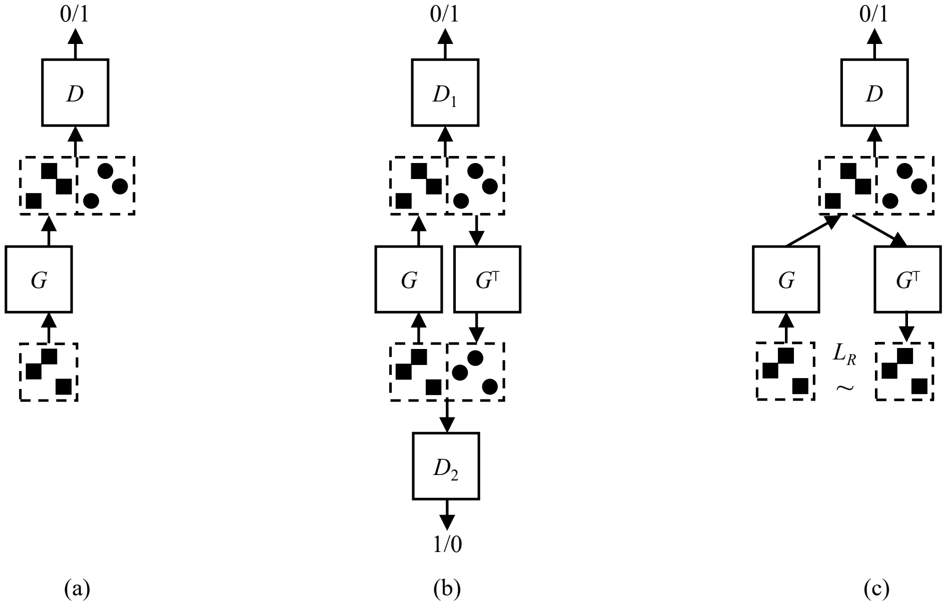

Figure 2: (a) The unidirectional transformation model directly inspired by the adversarial game: The generator G tries to transform source word embeddings (squares) to make them seem like target ones (dots), while the discriminator D tries to classify whether the input embeddings are generated by G or real samples from the target embedding distribution. (b) The bidirectional transformation model. Two generators with tied weights perform transformation between languages. Two separate discriminators are responsible for each language. (c) The adversarial autoencoder model. The generator aims to make the transformed embeddings not only indistinguishable by the discriminator, but also recoverable as measured by the reconstruction loss L_R.

because data encoding cross-lingual equivalence is often expensive to obtain.

In this work, we aim to entirely eliminate the need for cross-lingual supervision. Our approach draws inspiration from recent advances in generative adversarial networks (Goodfellow et al., 2014). We first formulate our task in a fashion that naturally admits an adversarial game. Then we propose three models that implement the game, and explore techniques to ensure the success of training. Finally, our evaluation on the bilingual lexicon induction task reveals encouraging performance, even though this task appears formidable without any cross-lingual supervision.

2 Models

In order to induce a bilingual lexicon, we start from two sets of monolingual word embeddings with dimensionality d. They are trained separately on two languages. Our goal is to learn a mapping function $f : \mathbb{R}^d \to \mathbb{R}^d$ so that for a source word embedding x, $f(x)$ lies close to the embedding of its target language translation y. The learned mapping function can then be used to translate each source word x by finding the nearest target embedding to $f(x)$.

We consider x to be drawn from a distribution p_x, and similarly $y \sim p_y$. The key intuition here is to find the mapping function to make $f(x)$ seem to follow the distribution p_y, for all $x \sim p_x$. From this point of view, we design an adversarial game as illustrated in Figure 2(a): The generator G implements the mapping function f, trying to make $f(x)$ passable as target word embeddings, while the discriminator D is a binary classifier striving to distinguish between fake target word embeddings $f(x) \sim p_{f(x)}$ and real ones $y \sim p_y$. This intuition can be formalized as the minimax game $\min_G \max_D V(D, G)$ with value function

$$\begin{aligned} & V(D, G) \\ =& \mathbb{E}_{y \sim p_y}[\log D(y)] + \\ & \mathbb{E}_{x \sim p_x}[\log(1 - D(G(x)))]. \end{aligned} \tag{1}$$

Theoretical analysis reveals that adversarial training tries to minimize the Jensen-Shannon divergence JSD $(p_y || p_{f(x)})$ (Goodfellow et al., 2014). Importantly, the minimization happens at the distribution level, without requiring word translation pairs to supervise training.

2.1 Model 1: Unidirectional Transformation

The first model directly implements the adversarial game, as shown in Figure 2(a). As hinted by the isomorphism shown in Figure 1, previous works typically choose the mapping function f to be a linear map (Mikolov et al., 2013a; Dinu et al., 2015; Lazaridou et al., 2015). We therefore parametrize the generator as a transformation matrix $G \in \mathbb{R}^{d \times d}$. We also tried non-linear maps parametrized by neural networks, without success. In fact, if the generator is given sufficient capacity, it can in principle learn a constant mapping function to a target word embedding, which makes the discriminator impossible to distinguish, much like the "mode collapse" problem widely observed in the image domain (Radford et al., 2015; Salimans et al., 2016). We therefore believe it is crucial to grant the generator with suitable capacity.

As a generic binary classifier, a standard feed-forward neural network with one hidden layer is used to parametrize the discriminator D, and its loss function is the usual cross-entropy loss, as in the value function (1):

$$L_D = -\log D\left(y\right) - \log\left(1 - D\left(Gx\right)\right). \quad (2)$$

For simplicity, here we write the loss with a mini-batch size of 1; in our experiments we use 128.

The generator loss is given by

$$L_G = -\log D\left(Gx\right). \quad (3)$$

In line with previous work (Goodfellow et al., 2014), we find this loss easier to minimize than the original form $\log\left(1 - D\left(Gx\right)\right)$.

Orthogonal Constraint

The above model is very difficult to train. One possible reason is that the parameter search space $\mathbb{R}^{d \times d}$ for the generator may still be too large. Previous works have attempted to constrain the transformation matrix to be orthogonal (Xing et al., 2015; Zhang et al., 2016b; Artetxe et al., 2016). An orthogonal transformation is also theoretically appealing for its self-consistency (Smith et al., 2017) and numerical stability. However, using constrained optimization for our purpose is cumbersome, so we opt for an orthogonal parametrization (Mhammedi et al., 2016) of the generator instead.

2.2 Model 2: Bidirectional Transformation

The orthogonal parametrization is still quite slow. We can relax the orthogonal constraint and only require the transformation to be self-consistent (Smith et al., 2017): If G transforms the source word embedding space into the target language space, its transpose $G^\top$ should transform the target language space back to the source. This can be implemented by two unidirectional models with a tied generator, as illustrated in Figure 2(b). Two separate discriminators are used, with the same cross-entropy loss as Equation (2) used by Model 1. The generator loss is given by

$$L_G = -\log D_1\left(Gx\right) - \log D_2\left(G^\top x\right). \quad (4)$$

2.3 Model 3: Adversarial Autoencoder

As another way to relax the orthogonal constraint, we introduce the adversarial autoencoder (Makhzani et al., 2015), depicted in Figure 2(c). After the generator G transforms a source word embedding x into a target language representation Gx, we should be able to reconstruct the source word embedding x by mapping back with $G^\top$. We therefore introduce the reconstruction loss measured by cosine similarity:

$$L_R = -\cos\left(x, G^\top Gx\right). \quad (5)$$

Note that this loss will be minimized if G is orthogonal. With this term included, the loss function for the generator becomes

$$L_G = -\log D\left(Gx\right) - \lambda \cos\left(x, G^\top Gx\right), \quad (6)$$

where λ is a hyperparameter that balances the two terms. $\lambda = 0$ recovers the unidirectional transformation model, while larger λ should enforce a stricter orthogonal constraint.

3 Training Techniques

Generative adversarial networks are notoriously difficult to train, and investigation into stabler training remains a research frontier (Radford et al., 2015; Salimans et al., 2016; Arjovsky and Bottou, 2017). We contribute in this aspect by reporting techniques that are crucial to successful training for our task.

3.1 Regularizing the Discriminator

Recently, it has been suggested to inject noise into the input to the discriminator (Sønderby et al., 2016; Arjovsky and Bottou, 2017). The noise is typically additive Gaussian. Here we explore more possibilities, with the following types of noise, injected into the input and hidden layer:

- Multiplicative Bernoulli noise (dropout) (Srivastava et al., 2014): $\epsilon \sim \text{Bernoulli}(p)$.
- Additive Gaussian noise: $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$.
- Multiplicative Gaussian noise: $\epsilon \sim \mathcal{N}\left(1, \sigma^2\right)$.

As noise injection is a form of regularization (Bishop, 1995; Van der Maaten et al., 2013; Wager et al., 2013), we also try l_2 regularization, and directly restricting the hidden layer size to combat overfitting. Our findings include:

- Without regularization, it is not impossible for the optimizer to find a satisfactory parameter configuration, but the hidden layer size has to be tuned carefully. This indicates that a balance of capacity between the generator and discriminator is needed.
- All forms of regularization help training by allowing us to liberally set the hidden layer size to a relatively large value.
- Among the types of regularization, multiplicative Gaussian injected into the input is the most effective, and additive Gaussian is similar. On top of input noise, hidden layer noise helps slightly.

In the following experiments, we inject multiplicative Gaussian into the input and hidden layer of the discriminator with $\sigma = 0.5$.

3.2 Model Selection

From a typical training trajectory shown in Figure 3, we observe that training is not convergent. In fact, simply using the model saved at the end of training gives poor performance. Therefore we need a mechanism to select a good model. We observe there are sharp drops of the generator loss L_G, and find they correspond to good models, as the discriminator gets confused at these points with its classification accuracy (D accuracy) dropping simultaneously. Interestingly, the reconstruction loss L_R and the value of $\left\|G^\top G - I\right\|_F$ exhibit synchronous drops, even if we use the unidirectional transformation model ($\lambda = 0$). This means a good transformation matrix is indeed nearly orthogonal, and justifies our encouragement of G towards orthogonality. With this finding, we can train for sufficient steps and save the model with the lowest generator loss.

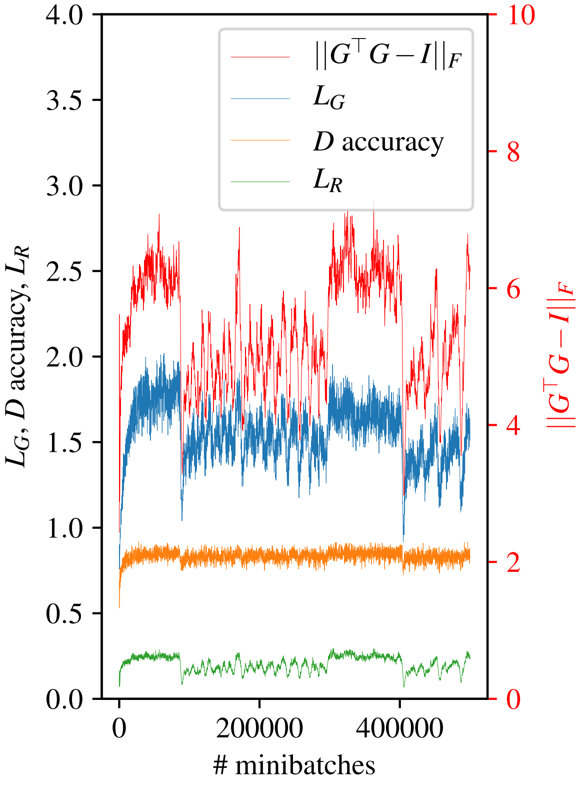

Figure 3: A typical training trajectory of the adversarial autoencoder model with $\lambda = 1$. The values are averages within each minibatch.

As we aim to find the cross-lingual transformation without supervision, it would be ideal to determine hyperparameters without a validation set. The sharp drops can also be indicative in this case. If a hyperparameter configuration is poor, those values will oscillate without a clear drop. Although this criterion is somewhat subjective, we find it to be quite feasible in practice.

3.3 Other Training Details

Our approach takes monolingual word embeddings as input. We train the CBOW model (Mikolov et al., 2013b) with default hyperparameters in `word2vec`.[1] The embedding dimension d is 50 unless stated otherwise. Before feeding them into our system, we normalize the word embeddings to unit length. When sampling words for adversarial training, we penalize frequent words in a way similar to (Mikolov et al., 2013b). G is

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

initialized with a random orthogonal matrix. The hidden layer size of D is 500. Adversarial training involves alternate gradient update of the generator and discriminator, which we implement with a simpler variant algorithm described in (Nowozin et al., 2016). Adam (Kingma and Ba, 2014) is used as the optimizer, with default hyperparameters. For the adversarial autoencoder model, $\lambda = 1$ generally works well, but $\lambda = 10$ appears stabler for the low-resource Turkish-English setting.

4 Experiments

We evaluate the quality of the cross-lingual embedding transformation on the bilingual lexicon induction task. After a source word embedding is transformed into the target space, its M nearest target embeddings (in terms of cosine similarity) are retrieved, and compared against the entry in a ground truth bilingual lexicon. Performance is measured by top-M accuracy (Vulić and Moens, 2013): If any of the M translations is found in the ground truth bilingual lexicon, the source word is considered to be handled correctly, and the accuracy is calculated as the percentage of correctly translated source words. We generally report the harshest top-1 accuracy, unless when comparing with published figures in Section 4.4.

Baselines

Almost all approaches to bilingual lexicon induction from non-parallel data depend on seed lexica. An exception is decipherment (Dou and Knight, 2012; Dou et al., 2015), and we use it as our baseline. The decipherment approach is not based on distributional semantics, but rather views the source language as a cipher for the target language, and attempts to learn a statistical model to decipher the source language. We run the MonoGiza system as recommended by the toolkit.[2] It can also utilize monolingual embeddings (Dou et al., 2015); in this case, we use the same embeddings as the input to our approach.

Sharing the underlying spirit with our approach, related methods also build upon monolingual word embeddings and find transformation to link different languages. Although they need seed word translation pairs to train and thus not directly comparable, we report their performance with 50 and 100 seeds for reference. These methods are:

		# tokens	vocab. size
Wikipedia comparable corpora			
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
Large-scale settings			
zh-en Wikipedia	zh	143m	14,686
	en	1,907m	61,899
zh-en Gigaword	zh	2,148m	45,958
	en	5,017m	73,504

Table 1: Statistics of the non-parallel corpora. Language codes: zh = Chinese, en = English, es = Spanish, it = Italian, ja = Japanese, tr = Turkish.

- Translation matrix (`TM`) (Mikolov et al., 2013a): the pioneer of this type of methods mentioned in the introduction, using linear transformation. We use a publicly available implementation.[3]

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

We ensure the same input embeddings for these methods and ours.

The seed word translation pairs are obtained as follows. First, we ask Google Translate[4] 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.

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

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

[4] https://translate.google.com

method	# seeds	accuracy (%)
MonoGiza w/o emb.	0	0.05
MonoGiza w/ emb.	0	0.09
TM	50	0.29
	100	21.79
IA	50	18.71
	100	32.29
Model 1	0	39.25
Model 1 + ortho.	0	28.62
Model 2	0	40.28
Model 3	0	43.31

Table 2: Chinese-English top-1 accuracies of the MonoGiza baseline and our models, along with the translation matrix (TM) and isometric alignment (IA) methods that utilize 50 and 100 seeds.

4.1 Experiments on Chinese-English

Data

For this set of experiments, the data for training word embeddings comes from Wikipedia comparable corpora.[5] Following (Vulić and Moens, 2013), we retain only nouns with at least 1,000 occurrences. For the Chinese side, we first use OpenCC[6] to normalize characters to be simplified, and then perform Chinese word segmentation and POS tagging with THULAC.[7] The preprocessing of the English side involves tokenization, POS tagging, lemmatization, and lowercasing, which we carry out with the NLTK toolkit.[8] The statistics of the final training data is given in Table 1, along with the other experimental settings.

As the ground truth bilingual lexicon for evaluation, we use Chinese-English Translation Lexicon Version 3.0 (LDC2002L27).

Overall Performance

Table 2 lists the performance of the MonoGiza baseline and our four variants of adversarial training. MonoGiza obtains low performance, likely due to the harsh evaluation protocol (cf. Section 4.4). Providing it with syntactic information can help (Dou and Knight, 2013), but in a low-resource scenario with zero cross-lingual information, parsers are likely to be inaccurate or even unavailable.

城市 *chengshi*	小行星 *xiaoxingxing*	文学 *wenxue*
city	**asteroid**	poetry
town	astronomer	**literature**
suburb	comet	prose
area	constellation	poet
proximity	orbit	writing

Table 3: Top-5 English translation candidates proposed by our approach for some Chinese words. The ground truth is marked in bold.

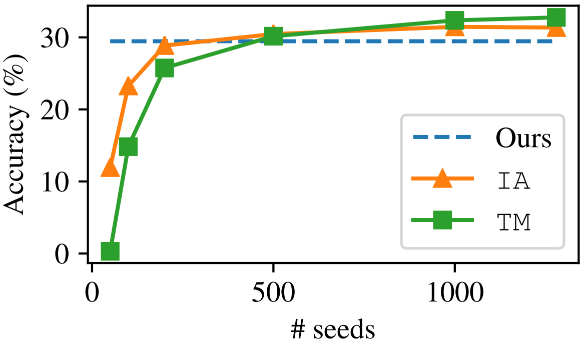

Figure 4: Top-1 accuracies of our approach, isometric alignment (IA), and translation matrix (TM), with the number of seeds varying in {50, 100, 200, 500, 1000, 1280}.

The unidirectional transformation model attains reasonable accuracy if trained successfully, but it is rather sensitive to hyperparameters and initialization. This training difficulty motivates our orthogonal constraint. But imposing a strict orthogonal constraint hurts performance. It is also about 20 times slower even though we utilize orthogonal parametrization instead of constrained optimization. The last two models represent different relaxations of the orthogonal constraint, and the adversarial autoencoder model achieves the best performance. We therefore use it in our following experiments. Table 3 lists some word translation examples given by the adversarial autoencoder model.

Comparison With Seed-Based Methods

In this section, we investigate how many seeds TM and IA require to attain the performance level of our approach. There are a total of 1,280 seed translation pairs for Chinese-English, which are removed from the test set during the evaluation for this experiment. We use the most frequent S pairs for TM and IA.

Figure 4 shows the accuracies with respect to

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

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

[7] http://thulac.thunlp.org

[8] http://www.nltk.org

method	# seeds	es-en	it-en	ja-zh	tr-en
MonoGiza w/o embeddings	0	0.35	0.30	0.04	0.00
MonoGiza w/ embeddings	0	1.19	0.27	0.23	0.09
TM	50	1.24	0.76	0.35	0.09
	100	48.61	37.95	26.67	11.15
IA	50	39.89	27.03	19.04	7.58
	100	60.44	46.52	36.35	17.11
Ours	0	71.97	58.60	43.02	17.18

Table 4: Top-1 accuracies (%) of the MonoGiza baseline and our approach on Spanish-English, Italian-English, Japanese-Chinese, and Turkish-English. The results for translation matrix (TM) and isometric alignment (IA) using 50 and 100 seeds are also listed.

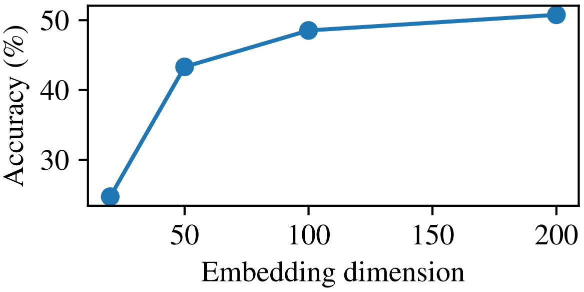

Figure 5: Top-1 accuracies of our approach with respect to the input embedding dimensions in {20, 50, 100, 200}.

S. When the seeds are few, the seed-based methods exhibit clear performance degradation. In this case, we also observe the importance of the orthogonal constraint from the superiority of IA to TM, which supports our introduction of this constraint as we attempt zero supervision. Finally, in line with the finding in (Vulić and Korhonen, 2016), hundreds of seeds are needed for TM to generalize. Only then do seed-based methods catch up with our approach, and the performance difference is marginal even when more seeds are provided.

Effect of Embedding Dimension

As our approach takes monolingual word embeddings as input, it is conceivable that their quality significantly affects how well the two spaces can be connected by a linear map. We look into this aspect by varying the embedding dimension d in Figure 5. As the dimension increases, the accuracy improves and gradually levels off. This indicates that too low a dimension hampers the encoding of linguistic information drawn from the corpus, and it is advisable to use a sufficiently large dimension.

4.2 Experiments on Other Language Pairs

Data

We also induce bilingual lexica from Wikipedia comparable corpora for the following language pairs: Spanish-English, Italian-English, Japanese-Chinese, and Turkish-English. For Spanish-English and Italian-English, we choose to use TreeTagger[9] for preprocessing, as in (Vulić and Moens, 2013). For the Japanese corpus, we use MeCab[10] 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. Unlike the other language pairs, the frequency cutoff threshold for Turkish-English is 100, as the amount of data is relatively small.

The ground truth bilingual lexica for Spanish-English and Italian-English are obtained from Open Multilingual WordNet[11] 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 way as how we obtain seed word translation pairs from Google Translate, described above.

Results

As shown in Table 4, the MonoGiza baseline still does not work well on these language pairs, while our approach achieves much better performance. The accuracies are particularly high for Spanish-English and Italian-English, likely because they are closely related languages, and their embedding spaces may exhibit stronger isomorphism. The

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

[10] http://taku910.github.io/mecab

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

method	# seeds	Wikipedia	Gigaword
TM	50	0.00	0.01
	100	4.79	2.07
IA	50	3.25	1.68
	100	7.08	4.18
Ours	0	7.92	2.53

Table 5: Top-1 accuracies (%) of our approach to inducing bilingual lexica for Chinese-English from Wikipedia and Gigaword. Also listed are results for translation matrix (TM) and isometric alignment (IA) using 50 and 100 seeds.

performance on Japanese-Chinese is lower, on a comparable level with Chinese-English (cf. Table 2), and these languages are relatively distantly related. Turkish-English represents a low-resource scenario, and therefore the lexical semantic structure may be insufficiently captured by the embeddings. The agglutinative nature of Turkish can also add to the challenge.

4.3 Large-Scale Settings

We experiment with large-scale Chinese-English data from two sources: the whole Wikipedia dump and Gigaword (LDC2011T13 and LDC2011T07). We also simplify preprocessing by removing the noun restriction and the lemmatization step (cf. preprocessing decisions for the above experiments).

Although large-scale data may benefit the training of embeddings, it poses a greater challenge to bilingual lexicon induction. First, the degree of non-parallelism tends to increase. Second, with cruder preprocessing, the noise in the corpora may take its toll. Finally, but probably most importantly, the vocabularies expand dramatically compared to previous settings (see Table 1). This means a word translation has to be retrieved from a much larger pool of candidates.

For these reasons, we consider the performance of our approach presented in Table 5 to be encouraging. The imbalanced sizes of the Chinese and English Wikipedia do not seem to cause a problem for the structural isomorphism needed by our method. MonoGiza does not scale to such large vocabularies, as it already takes days to train in our Italian-English setting. In contrast, our approach is immune from scalability issues by working with embeddings provided by word2vec, which is well known for its fast speed. With the network configuration used in our experiments, the adversarial autoencoder model takes about two hours to train for 500k minibatches on a single CPU.

method	5k	10k
MonoGiza w/o embeddings	13.74	7.80
MonoGiza w/ embeddings	17.98	10.56
(Cao et al., 2016)	23.54	17.82
Ours	68.59	51.86

Table 6: Top-5 accuracies (%) of 5k and 10k most frequent words in the French-English setting. The figures for the baselines are taken from (Cao et al., 2016).

4.4 Comparison With (Cao et al., 2016)

In order to compare with the recent method by Cao et al. (2016), which also uses zero cross-lingual signal to connect monolingual embeddings, we replicate their French-English experiment to test our approach.[12] This experimental setting has important differences from the above ones, mostly in the evaluation protocol. Apart from using top-5 accuracy as the evaluation metric, the ground truth bilingual lexicon is obtained by performing word alignment on a parallel corpus. We find this automatically constructed bilingual lexicon to be noisier than the ones we use for the other language pairs; it often lists tens of translations for a source word. This lenient evaluation protocol should explain MonoGiza's higher numbers in Table 6 than what we report in the other experiments. In this setting, our approach is able to considerably outperform both MonoGiza and the method by Cao et al. (2016).

5 Related Work

5.1 Cross-Lingual Word Embeddings for Bilingual Lexicon Induction

Inducing bilingual lexica from non-parallel data is a long-standing cross-lingual task. Except for the decipherment approach, traditional statistical methods all require cross-lingual signals (Rapp, 1999; Koehn and Knight, 2002; Fung and Cheung, 2004; Gaussier et al., 2004; Haghighi et al., 2008; Vulić et al., 2011; Vulić and Moens, 2013).

Recent advances in cross-lingual word embeddings (Vulić and Korhonen, 2016; Upadhyay et al.,

[12] As a confirmation, we ran MonoGiza in this setting and obtained comparable performance as reported.

2016) have rekindled interest in bilingual lexicon induction. Like their traditional counterparts, these embedding-based methods require cross-lingual signals encoded in parallel data, aligned at document level (Vulić and Moens, 2015), 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 word level (i.e. seed lexicon) (Gouws and Søgaard, 2015; Wick et al., 2016; Duong et al., 2016; Shi et al., 2015; Mikolov et al., 2013a; Dinu et al., 2015; Lazaridou et al., 2015; Faruqui and Dyer, 2014; Lu et al., 2015; Ammar et al., 2016; Zhang et al., 2016a, 2017; Smith et al., 2017). In contrast, our work completely removes the need for cross-lingual signals to connect monolingual word embeddings, trained on non-parallel text corpora.

As one of our baselines, the method by Cao et al. (2016) also does not require cross-lingual signals to train bilingual word embeddings. It modifies the objective for training embeddings, whereas our approach uses monolingual embeddings trained beforehand and held fixed. More importantly, its learning mechanism is substantially different from ours. It encourages word embeddings from different languages to lie in the shared semantic space by matching the mean and variance of the hidden states, assumed to follow a Gaussian distribution, which is hard to justify. Our approach does not make any assumptions and directly matches the mapped source embedding distribution with the target distribution by adversarial training.

A recent work also attempts adversarial training for cross-lingual embedding transformation (Barone, 2016). The model architectures are similar to ours, but the reported results are not positive. We tried the publicly available code on our data, but the results were not positive, either. Therefore, we attribute the outcome to the difference in the loss and training techniques, but not the model architectures or data.

5.2 Adversarial Training

Generative adversarial networks are originally proposed for generating realistic images as an implicit generative model, but the adversarial training technique for matching distributions is generalizable to much more tasks, including natural language processing. For example, Ganin et al. (2016) address domain adaptation by adversarially training features to be domain invariant, and test on sentiment classification. Chen et al. (2016) extend this idea to cross-lingual sentiment classification. Our research deals with unsupervised bilingual lexicon induction based on word embeddings, and therefore works with word embedding distributions, which are more interpretable than the neural feature space of classifiers in the above works.

In the field of neural machine translation, a recent work (He et al., 2016) proposes dual learning, which also involves a two-agent game and therefore bears conceptual resemblance to the adversarial training idea. The framework is carried out with reinforcement learning, and thus differs greatly in implementation from adversarial training.

6 Conclusion

In this work, we demonstrate the feasibility of connecting word embeddings of different languages without any cross-lingual signal. This is achieved by matching the distributions of the transformed source language embeddings and target ones via adversarial training. The success of our approach signifies the existence of universal lexical semantic structure across languages. Our work also opens up opportunities for the processing of extremely low-resource languages and domains that lack parallel data completely.

Our work is likely to benefit from advances in techniques that further stabilize adversarial training. Future work also includes investigating other divergences that adversarial training can minimize (Nowozin et al., 2016), and broader mathematical tools that match distributions (Mohamed and Lakshminarayanan, 2016).

Acknowledgments

We thank the anonymous reviewers for their helpful 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.

References

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

Martin Arjovsky and Léon Bottou. 2017. Towards Principled Methods For Training Generative Adversarial Networks. In *ICLR*. http://arxiv.org/abs/1701.04862.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*. http://aclanthology.info/papers/learning-principled-bilingual-mappings-of-word-embeddings-while-preserving-monolingual-invariance.

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*. https://doi.org/10.18653/v1/W16-1614.

Chris M. Bishop. 1995. Training with Noise is Equivalent to Tikhonov Regularization. *Neural Comput.* https://doi.org/10.1162/neco.1995.7.1.108.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A Distribution-based Model to Learn Bilingual Word Embeddings. In *COLING*. http://aclanthology.info/papers/a-distribution-based-model-to-learn-bilingual-word-embeddings.

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*. http://papers.nips.cc/paper/5270-an-autoencoder-approach-to-learning-bilingual-word-representations.pdf.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. *arXiv:1606.01614 [cs]* http://arxiv.org/abs/1606.01614.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, Fast Cross-lingual Word-embeddings. In *EMNLP*. http://aclanthology.info/papers/trans-gram-fast-cross-lingual-word-embeddings.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *ACL-HLT*. http://aclweb.org/anthology/P11-2071.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. In *ICLR Workshop*. http://arxiv.org/abs/1412.6568.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *EMNLP-CoNLL*. http://aclweb.org/anthology/D12-1025.

Qing Dou and Kevin Knight. 2013. Dependency-Based Decipherment for Resource-Limited Machine Translation. In *EMNLP*. http://aclanthology.info/papers/dependency-based-decipherment-for-resource-limited-machine-translation.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying Bayesian Inference and Vector Space Models for Improved Decipherment. In *ACL-IJCNLP*. http://www.aclweb.org/anthology/P15-1081.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning Crosslingual Word Embeddings without Bilingual Corpora. In *EMNLP*. http://aclanthology.info/papers/learning-crosslingual-word-embeddings-without-bilingual-corpora.

Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *EACL*. http://aclanthology.info/papers/improving-vector-space-word-representations-using-multilingual-correlation.

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *EMNLP*. http://aclweb.org/anthology/W04-3208.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* http://jmlr.org/papers/v17/15-239.html.

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*. https://doi.org/10.3115/1218955.1219022.

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*. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *ICML*. http://jmlr.org/proceedings/papers/v37/gouws15.html.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *NAACL-HLT*. http://www.aclweb.org/anthology/N15-1157.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *ACL-HLT*. http://aclanthology.info/papers/learning-bilingual-lexicons-from-monolingual-corpora.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual Learning for Machine Translation. In *NIPS*. http://papers.nips.cc/paper/6469-dual-learning-for-machine-translation.pdf.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Distributed Representations without Word Alignment. In *ICLR*. http://arxiv.org/abs/1312.6173.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. http://aclweb.org/anthology/W13-2233.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* http://arxiv.org/abs/1412.6980.

Philipp Koehn and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*. https://doi.org/10.3115/1118627.1118629.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning Bilingual Word Representations by Marginalizing Alignments. In *ACL*. http://aclanthology.info/papers/learning-bilingual-word-representations-by-marginalizing-alignments.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. In *ACL-IJCNLP*. https://doi.org/10.3115/v1/P15-1027.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep Multilingual Correlation for Improved Word Embeddings. In *NAACL-HLT*. http://aclanthology.info/papers/deep-multilingual-correlation-for-improved-word-embeddings.

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*. http://aclanthology.info/papers/bilingual-word-representations-with-monolingual-quality-in-mind.

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

Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. 2016. Efficient Orthogonal Parametrisation of Recurrent Neural Networks Using Householder Reflections. *arXiv:1612.00188 [cs]* http://arxiv.org/abs/1612.00188.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Shakir Mohamed and Balaji Lakshminarayanan. 2016. Learning in Implicit Generative Models. *arXiv:1610.03483 [cs, stat]* http://arxiv.org/abs/1610.03483.

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

Takamasa Oshikiri, Kazuki Fukui, and Hidetoshi Shimodaira. 2016. Cross-Lingual Word Representations via Spectral Graph Embeddings. In *ACL*. https://doi.org/10.18653/v1/P16-2080.

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

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL*. https://doi.org/10.3115/1034678.1034756.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *NIPS*. http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning Cross-lingual Word Embeddings via Matrix Co-factorization. In *ACL-IJCNLP*. http://aclanthology.info/papers/learning-cross-lingual-word-embeddings-via-matrix-co-factorization.

Samuel Smith, David Turban, Steven Hamblin, and Nils Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR*. http://arxiv.org/abs/1702.03859.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. 2016. Amortised MAP Inference for Image Super-resolution. *arXiv:1610.04490 [cs, stat]* http://arxiv.org/abs/1610.04490.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* http://www.jmlr.org/papers/v15/srivastava14a.html.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In *LREC*. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1138_Paper.pdf.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *ACL*. http://aclanthology.info/papers/cross-lingual-models-of-word-embeddings-an-empirical-comparison.

Laurens Van der Maaten, Minmin Chen, Stephen Tyree, and Kilian Weinberger. 2013. Learning with Marginalized Corrupted Features. In *ICML*. http://www.jmlr.org/proceedings/papers/v28/vandermaaten13.html.

Ivan Vulić and Anna Korhonen. 2016. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *ACL*. http://aclanthology.info/papers/on-the-role-of-seed-lexicons-in-learning-bilingual-word-embeddings.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses. In *NAACL-HLT*. http://aclanthology.info/papers/cross-lingual-semantic-similarity-of-words-as-the-similarity-of-their-semantic-word-responses.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In *ACL-IJCNLP*. http://aclanthology.info/papers/bilingual-word-embeddings-from-non-parallel-document-aligned-data-applied-to-bilingual-lexicon-induction.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying Word Translations from Comparable Corpora Using Latent Topic Models. In *ACL-HLT*. http://aclanthology.info/papers/identifying-word-translations-from-comparable-corpora-using-latent-topic-models.

Stefan Wager, Sida Wang, and Percy S Liang. 2013. Dropout Training as Adaptive Regularization. In *NIPS*. http://papers.nips.cc/paper/4882-dropout-training-as-adaptive-regularization.pdf.

Michael Wick, Pallika Kanani, and Adam Pocock. 2016. Minimally-Constrained Multilingual Embeddings via Artificial Code-Switching. In *AAAI*. http://www.aaai.org/Conferences/AAAI/2016/Papers/15Wick12464.pdf.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *NAACL-HLT*. http://aclanthology.info/papers/normalized-word-embedding-and-orthogonal-transform-for-bilingual-word-translation.

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* https://doi.org/10.1073/pnas.1520752113.

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*. http://aclanthology.info/papers/inducing-bilingual-lexica-from-non-parallel-data-with-earth-mover-s-distance-regularization.

Meng Zhang, Haoruo Peng, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Bilingual Lexicon Induction From Non-Parallel Data With Minimal Supervision. In *AAAI*. http://thunlp.org/~zm/publications/aaai2017.pdf.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016b. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *NAACL-HLT*. http://aclanthology.info/papers/ten-pairs-to-tag-multilingual-pos-tagging-via-coarse-mapping-between-embeddings.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP*. http://aclanthology.info/papers/bilingual-word-embeddings-for-phrase-based-machine-translation.