Bilingual Adversarial Autoencoder For Unsupervised Cross-lingual Word Representation Learning

Anonymous EMNLP submission

Abstract

Learning cross-lingual word representations (**CLWR**) in a joint embedding space typically requires cross-lingual signals as supervision. Recent works went further to learn CLWR without any form of supervision. However, this is achieved by transforming the source embeddings to fit in the target embedding space, which is pre-trained and therefore might be suboptimal for the source language. To address this issue, we propose a novel Bilingual Adversarial AutoEncoder (BiAAE) for unsupervised CLWR learning. Our approach jointly transforms the source and target embeddings into a new common embedding space where the embeddings of the two languages can be matched much better. By conducting extensive experiments on 6 language pairs, we demonstrate that the proposed method significantly outperforms current state-of-the-art unsupervised models.

1 Introduction

Learning Cross-lingual word representations (**CLWR**) in a joint common space has attracted a lot of attention in recent years. CLWR plays a very important role in many NLP tasks (Tsai and Roth, 2016; Upadhyay et al., 2016; Artetxe et al., 2017). Generally, CLWR can be studied with a transformation-based framework and an interlingua-based framework. In the transformation framework, CLWR are learned by transforming the embeddings from source space into a fixed and pre-trained target space (Mikolov et al., 2013b; Xing et al., 2015; Lazaridou et al., 2015; Artetxe et al., 2016). The interlingua-based framework represents words from different languages in a common embedding space to learn cross-lingual word embeddings (Hermann and Blunsom, 2014; Chandar et al., 2014; Faruqui and Dyer, 2014; Gouws et al., 2015; Coulmance et al., 2015; Luong et al., 2015). In spite of their success, all of these approaches require supervision from cross-lingual signals. This is unfortunate for low-resource languages and domains.

To address this issue, researchers have been initiated to explore the transformation-based framework for unsupervised CLWR learning (Zhang et al., 2017a,b; Conneau et al., 2017), which eliminate the requirement of bilingual parallel corpora. Typically, these methods are based on Generative Adversarial Networks (**GANs**) (Goodfellow et al., 2014), by which the embeddings from source vector space are transformed into the embedding space of the target space, to learn cross-lingual word embeddings and show encouraging performance. However, the transformation-based framework relies heavily on a pre-trained target language embedding space, of which cannot adequately capture the characteristics of cross-lingual, especially source language.

In this paper, we proposed an interlingua-based unsupervised framework to learn cross-lingual word representation. In particular, we leverage adversarial training to learn an autoencoder for each language to encode source and target word embeddings jointly into a common space, which not only preserves the monolingual invariance, but also cross-lingual characteristics. By conducting extensive experiments on 6 language pairs, we demonstrate that the proposed method significantly outperforms several state-of-the-art unsupervised models.

In summary, this paper makes the following contributions:

- We propose a novel Bilingual Adversarial AutoEncoder (BiAAE) for unsupervised cross-lingual word representation learning.

In particular, our model jointly learns the source and target language embeddings in a common space.

- Extensive experiments show that our model significantly outperforms state-of-the-art unsupervised models.

2 Background

In this section, we introduce unsupervised transformation-based framework and unsupervised transformation-based Models, which are the basics of this paper. Given the monolingual data of source and target language, let $X_1 \in \mathbb{R}^{d_{X_1}}$ and $X_2 \in \mathbb{R}^{d_{X_2}}$ denote two sets of embeddings trained on two monolingual data, respectively. d_{X_1} and d_{X_2} are vector dimension.

2.1 Transformation-based Framework

In the transformation-based framework, cross-lingual word embeddings are constructed by a mapping function f:$\mathbb{R}^{d_{X_1}} \rightarrow \mathbb{R}^{d_{X_2}}$, which transforms word embeddings X_1 of source language to target language embedding space X_2. Given the seed dictionary from bilingual corpus, the mapping function f can be learned by minimizing the distance between the transformed embedding $f(x_1)$ and the target embedding x_2 as:

$$\ell(X_1, X_2) = \mathbb{E}_{(\hat{x}_1, \hat{x}_2)}[\triangle(f(x_1), x_2)], \quad (1)$$

where $(\hat{x}_1, \hat{x}_2)$ is a translation pair uniformly sampled from the seed dictionary, $\triangle$ denotes the distance criterion (i.e., cosin). For the mapping function f, a linear transformation matrix is defined in the pioneer transformation-based framework (Mikolov et al., 2013b). Followed attempts (Xing et al., 2015; Zhang et al., 2016; Artetxe et al., 2016; Smith et al., 2017) extended this framework and showed that the results are improved by enforcing f as a orthogonal matrix. In this paper, we use a linear transformation matrix for better generalization.

2.2 Unsupervised Transformation-based Models

The traditional transformation-based framework is further extended to the unsupervised scenario by eliminating the requirement of the seed dictionary (Zhang et al., 2017a,b; Conneau et al., 2017). Specifically, the unsupervised transformation-based methods are to connecting word embeddings of the source and target language via adversarial training. Formally, they formalize this task as an adversarial game:

i) A discriminator D is introduced to classify the transformed embeddings $f(X_1)$ and the target embeddings X_2 by maximizing the following objective:

$$\ell_D(X_1, X_2) = \mathbb{E}_{\hat{x}_2 \sim X_2}[\log D(\hat{x}_2)] + \mathbb{E}_{\hat{x}_1 \sim X_1}[\log(1 - D_1(f(\hat{x}_1))], \quad (2)$$

where $\hat{x}_1 \sim X_1$ denotes that $\hat{x}_1$ is sampled from X_1, and $\hat{x}_2 \sim X_2$ denotes that $\hat{x}_2$ is sampled from X_2.

ii) The mapping function f is trained to confuse the discriminator by minimizing the following objective:

$$\ell_{adv}(X_1) = \mathbb{E}_{\hat{x}_1 \sim X_1}[\log(1 - D(f(\hat{x}_1))], \quad (3)$$

where f is a orthogonal matrix. The mapping f is then learned by minimizing the objective function:

$$\ell_{total} = \ell_{adv}(X_1) + \lambda \ell_R, \quad (4)$$

where ℓ_R is a regularization as the transformation-based framework. Finally, the mapping f and the discriminator D are learned by the adversarial training, thus connecting word embeddings of different languages without any cross-lingual signal.

3 Bilingual Adversarial Autoencoder

In this section, we first describe the intuition behind our method and then formally present the model.

3.1 Model Design

Compared with the existing unsupervised methods, we explore to transform the source and target embeddings to a new common embedding space rather than transforming the source embeddings to fit in the target embedding space. In other words, we try to jointly learn source and target cross-lingual embeddings in a common embedding space, where the embddings of source and target language can be matched better. To this end, two autoencoders are designed to model source mapping function $f_{x_1} : \mathbb{R}^{d_{X_1}} \rightarrow \mathbb{R}^{d_Z}$ and target mapping function $f_{x_2} : \mathbb{R}^{d_{X_2}} \rightarrow \mathbb{R}^{d_Z}$, and the adversarial strategy is used to construct (encode) cross-lingual word embeddings according.

Compared with monolingual word embedding X_1 and X_2, let $[X_{1Z}, X_{2Z}]$ denote the learned cross-lingual word embeddings. The $Encoder_1$ and $Decoder_1$ denote the encoder and decoder of source language, respectively. The $Encoder_2$ and $Decoder_2$ denote the encoder and decoder of target language, respectively. The D_1, D_2 denote two discriminators.

Intuitively, our motivations are illustrated as:

i) Self-reconstruction. We hope each autoencoder preserves the monolingual characteristics of source (or target) language. Specifically, the $Decoder_1$ (or $Decoder_2$) is used to reconstruct monolingual word embedding X_1 (or X_2) from the encoded cross-lingual word embedding X_{1Z} (or X_{2Z}), thus generating a self-reconstructed embeddings $\tilde{X}_1^{1\to1}$ (or $\tilde{X}_2^{2\to2}$).

ii) Cross-transformation. These learned embeddings are expected to capture cross-lingual characteristics. In particular, we hope the source cross-lingual embeddings X_{1Z} and target cross-lingual embeddings X_{2Z} share the same semantic space such that: 1) Cross-lingual word embeddings of target language X_{2Z} can be "cross-transformed" to source embedding space by the source language $Decoder_1$. 2) Cross-lingual word embeddings of source language X_{1Z} can be "cross-transformed" to target embedding space by the target language $Decoder_2$. To enforce this constraint, the discriminator D_1 for source language is used to classify between the cross-transformed embeddings $\tilde{X}_1^{2\to1}$ and source embeddings X_1. Similarly, the discriminator D_2 for target language is used to classify between the cross-transformed embeddings $\tilde{X}_2^{1\to2}$ and target embeddings X_2. The total architecture of our model is shown in Figure 1.

3.2 Model Architecture

Following the principles mentioned in Section 3.1, this section will introduce the architecture and training processing of the proposed model in detail.

Monolingual Auto-Encoding First, we hope that the induced cross-lingual word embeddings can preserve original monolingual information as much as possible, as done in standard auto-encoders (Rumelhart et al., 1986). Take source language as an example, an autoencoder $[Encoder_1, Decoder_1]$ is trained to encode word representations X_1 of source language. The encoder $Encoder_1$ is denoted as the function f_{x1} which maps the X_1 into X_{1Z} (corresponding to the mapping mentioned in 3.1). A decoder function g_{x1} is then used to get the self-reconstruction embedding $\tilde{X}_1^{1\to1}$. In this work, both of f_{x1} and g_{x1} are defined as linear mapping functions following the idea of Mikolov et al. (2013b), respectively.

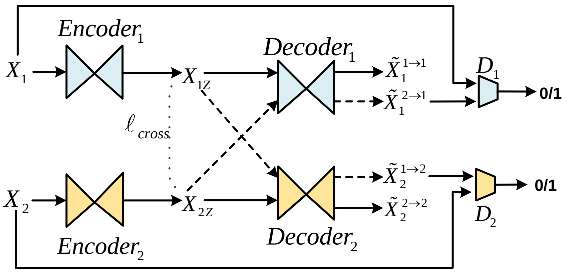

Figure 1: The proposed BiAAE framework. Two autoencoders $[Encoder_1, Decoder_1]$ and $[Encoder_2, Decoder_2]$ are jointly trained to induce cross-lingual word embeddings X_{1Z} and X_{2Z} from source word embeddings X_1 and target word embeddings X_2. Note that $\tilde{X}_1^{1\to1}$ and $\tilde{X}_2^{2\to2}$ are self-reconstructed embeddings; $\tilde{X}_1^{2\to1}$ and $\tilde{X}_2^{1\to2}$ are cross-transformed embeddings. D_1 and D_2 are discriminators to distinguish the cross-transformed embeddings and monolingual embeddings.

For word embeddings X_1 and X_2 from source and target language, their self-reconstructed embeddings are denoted as $\tilde{X}_1^{1\to1}$ and $\tilde{X}_2^{2\to2}$. Two autoencoders are trained to minimize a loss function which measures the discrepancy between original and self-reconstructed word embeddings as:

$$\ell_{mono}(X_1) = \mathbb{E}_{\hat{x}_1 \sim X_1}[\triangle(\hat{x}_1, \tilde{x}_1^{1\to1}] = \mathbb{E}_{\hat{x}_1 \sim X_1}[\triangle(\hat{x}_1, g_{x_1}(f_{x_1}(\hat{x}_1)))], \quad (5)$$

$$\ell_{mono}(X_2) = \mathbb{E}_{\hat{x}_2 \sim X_2}[\triangle(\hat{x}_2, \tilde{x}_2^{2\to2}] = \mathbb{E}_{\hat{x}_2 \sim X_2}[\triangle(\hat{x}_2, g_{x_2}(f_{x_2}(\hat{x}_2)))], \quad (6)$$

where $\hat{x} \sim X$ denotes that $\hat{x}$ is sampled from X, $\tilde{x}_1^{1\to1}$ is the self-reconstructed embedding of x_1, similarly $\tilde{x}_2^{2\to2}$ of x_2, $f_{x1}, g_{x1}, f_{x2}, g_{x2}$ is the encode and decode function of each autoencoder, and $\triangle$ denotes the discrepancy criterion, which is set as the average cosine similarity in our model.

Adversarial Training Recent methods have shown that the adversarial training can be used to match two distributions (Zhang et al., 2017a,b; Conneau et al., 2017; Lample et al., 2017).

Inspired by the above method, We connect the distribution of X_{1Z} and X_{2Z} by adversarial training. Specifically, When the following two constraints are satisfied at the same time, we consider the distribution of X_{1Z} and X_{2Z} are well connected: **Constraint 1**: Given encoded word embedding x_{1Z} from X_{1Z}, x_{1Z} can be transformed into the target language embedding space X_2 by $Decoder_2$; **Constraint 2**: Given encoded word embedding x_{2Z} from X_{2Z}, x_{2Z} can be transformed into the target language embedding space X_1 by $Decoder_1$.

Formally, let $\hat{x}_{1Z}$ be word embedding from the X_{1Z}, and $\hat{x}_{2Z}$ be word embedding from the X_{2Z}. We decode the x_{1Z} as $\tilde{x}_2^{1\to2}$, and decode the x_{2Z} as $\tilde{x}_1^{2\to1}$:

$$\begin{aligned} \tilde{x}_2^{1\to2} &= g_{x_2}(x_{1z}) = g_{x_2}(f_{x_1}(\hat{x}_1)), \\ \tilde{x}_1^{2\to1} &= g_{x_1}(x_{2z}) = g_{x_1}(f_{x_2}(\hat{x}_2)). \end{aligned} \tag{7}$$

We utilize adversarial training to satisfied the above constraint by matching the distribution of X_1 and the distribution of $\tilde{X}_1^{2\to1}$, and similarly matching the distribution of X_2 and the distribution of $\tilde{X}_2^{1\to2}$. To this end, the neural network-based discriminator D_1 is trained to classify the cross-transformed embeddings $\tilde{X}_1^{2\to1}$ and X_1, and similarly, discriminator D_2 for $\tilde{X}_2^{1\to2}$ and X_2. This intuition is formalized as following minimax game:

(1) Two discriminators are trained to classify the original distribution and the cross-transformed distribution by maximizing the following objective:

$$\begin{aligned} \ell_{D_1}(X_1, X_2) = \mathbb{E}_{\hat{x}_1\sim X_1}[\log D_1(\hat{x}_1)] + \\ \mathbb{E}_{\hat{x}_2\sim X_2}[\log(1 - D_1(g_{x_1}(f_{x_2}(\hat{x}_2))))], \end{aligned} \tag{8}$$

$$\begin{aligned} \ell_{D_2}(X_1, X_2) = \mathbb{E}_{\hat{x}_2\sim X_2}[\log D_2(\hat{x}_2)] + \\ \mathbb{E}_{\hat{x}_1\sim X_1}[\log(1 - D_2(g_{x_2}(f_{x_1}(\hat{x}_1))))]. \end{aligned} \tag{9}$$

(2) Two autoencoders are trained to confuse the discriminator by minimizing the following objective:

$$\begin{aligned} \ell_{adv}(X_2) = \\ \mathbb{E}_{\hat{x}_2\sim X_2}[\log(1 - D_1(g_{x_1}(f_{x_2}(\hat{x}_2))))], \end{aligned} \tag{10}$$

$$\begin{aligned} \ell_{adv}(X_1) = \\ \mathbb{E}_{\hat{x}_1\sim X_1}[\log(1 - D_2(g_{x_2}(f_{x_1}(\hat{x}_1))))]. \end{aligned} \tag{11}$$

Since the adversarial training happens at the distribution level, no cross-lingual supervision is needed.

Cross-lingual Constraint Note that there still exists "pseudo" cross-lingual word embedding pairs (x_{1Z}, x_{2Z}) which confuses the discriminator since the discriminator D_1 and D_2 could focus on a combination of local differences between the distributions, which is a common problem of GANs[1]. To tackle this issue, the following objective is proposed to constrain (x_{1Z}, x_{2Z}) be similar as much as possible:

$$\begin{aligned} \ell_{cross}(X_{1Z}, X_{2Z}) = \\ \mathbb{E}_{\hat{x}_1\sim X_1, \hat{x}_2\sim X_2}[\triangle(f_{x_1}(\hat{x}_1), f_{x_2}(\hat{x}_2))], \end{aligned} \tag{12}$$

where $\hat{x}, \hat{y}$ is sampled from X, Y separately, $\triangle$ denotes the criterion function, average cosine similarity again.

Total objective The total objective of autoencoders is:

$$\begin{aligned} \ell_{total} = \lambda_{cross}\ell_{cross}(X_{1Z}, X_{2Z}) + \\ \lambda_{mono}[\ell_{mono}(X_1) + \ell_{mono}(X_2)] + \\ \lambda_{adv}[\alpha\ell_{adv}(X_1) + (1-\alpha)\ell_{adv}(X_2)], \end{aligned} \tag{13}$$

where λ_{cross}, λ_{auto}, and λ_{adv} are weighting hyper-parameters for ℓ_{cross}, ℓ_{mono}, and ℓ_{adv}, respectively. Another hyper-parameter α is used to balance the importance of source and target autoencoders. At each iteration training, we train the autoencoders and discriminators in turn.

3.3 Model Selection

In order to select a best model, we wish to have a criterion correlated with the induced embedding quality. However, we do not have access to parallel data to judge how well our model works, not even at validation time. For the problem, Zhang et al. (2017a,b) propose to use "sharp drops" of generator loss to select model, while Conneau et al. (2017) propose CSLS as an unsupervised criterion which quantifies the closeness of produced cross-lingual word embeddings.[2] In this paper, we adopt CSLS to perform model selection. We first consider the 10k most frequent source words and use CSLS to generate a translation for each word, and compute the average cosine similarity between

[1] We found this problem occur in all language pairs in our experiment

[2] In our experiment, we found both criteria work well.

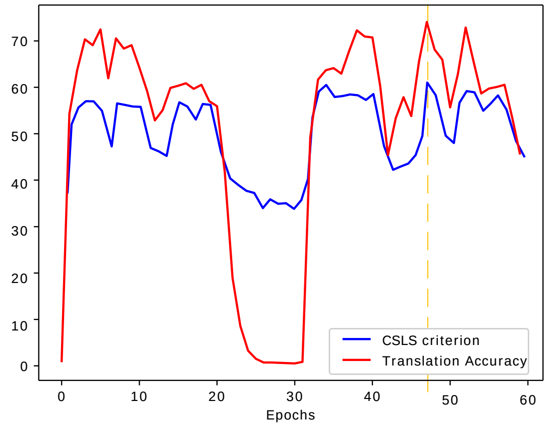

Figure 2: CSLS criterion and actual word translation accuracy in our experiment, we select a point where CSLS reaches to a maximum value (i.e., epoch 47).

these deemed translations. Finally, this average is used as a validation metric (also done by Conneau et al. (2017)). We show that the CSLS criterion is well correlated with the performance in Figure 2.

4 Experiments

In this section, we empirically measure the quality of the induced cross-lingual word embeddings in terms of their performance, when used as features in the following tasks: **(i)** Word Analogy; **(ii)** Word Translation; **(iii)** Sentence Translation Retrieval.

4.1 Datasets

We conduct our experiments on 6 language pairs: English-Italian (en-it), English-Spanish (en-es), English-French (en-fr), English-Russian (en-ru), English-Chinese (en-zh) and English-Esperanto (en-eo). To ensure comparability to previous methods, we used the code and datasets from the MUSE[3] repository by Conneau et al. (2017). They trained monolingual word vectors of dimension 300 with fastText[4] on Wikipedia monolingual corpora[5]. For English-Italian, to compare with other state-of-the-art methods, we use datasets released by Dinu and Baroni (2014)[6]. The word vectors were trained using word2vec, we use the Wacky/ukWaC and BNC corpora for English, while the Wacky/itWaC corpus for Italian. We select the first 200k most frequent words for evaluation in our experiments.

4.2 Evaluation Tasks

Word Analogy This task aims at evaluating the monolingual quality of the induced cross-lingual embeddings. We use the word analogy task proposed by (Mikolov et al., 2013a), which measures the accuracy on answering questions like “King - Man + Woman = ?” using simple word vector arithmetic. The dataset consists 8,869 semantic and 10,675 syntactic questions of this type, and is publicly available.[7] The vocabulary size is 30k (Mikolov et al., 2013a) With these settings, we obtain a coverage of 64.98%.

Word Translation This task judges the ability of cross-lingual embeddings to detect word pairs that are semantically similar across languages. We use the high-quality gold dictionaries released by Conneau et al. (2017) for evaluation. For each pair (e, f) in the gold dictionary, we check if f belongs to the list of top-k neighbors of e, according to the induced cross-lingual word vectors. For each language pair, we consider 1,500 query source and 200k target words. Precision@k for k in $\{1, 5\}$ are reported in this task.

Sentence Translation Retrieval In this task, we assess whether the learned cross-lingual representations are semantically coherent in sentence level across multiple languages. We follow Conneau et al. to use bag-of-words aggregation methods to perform sentence retrieval on the Europarl corpus. Sentence representations are computed by taking the tf-idf[8] weighted average of vectors of the words present in it. For each language pair, 2,000 source sentence queries and 200k target sentences are used for evaluation. The precision@k for k in $\{1, 5\}$ are reported, which judges whether the founded translation of the source sentence is in the top-k nearest neighbors.

4.3 Baselines

Although unsupervised CLRL is a very challenging problem, previous attempts have given strong baselines. We firstly compare our model with state of the art unsupervised models, and then we compare with the state of the art seed-based supervised models.

[3] https://github.com/facebookresearch/MUSE

[4] https://github.com/facebookresearch/fastText

[5] https://dumps.wikimedia.org/

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

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

[8] The idf weights are obtained using other 300k sentences from Europarl

Setting	en-es	en-fr	en-ru	en-zh
Mono	**80.31**	**80.31**	**80.31**	**80.31**
Zhang17a	79.45	80.21	79.47	79.50
Conneau17	80.22	80.31	80.05	79.67
BiAAE	80.16	79.43	78.76	79.54

Table 1: Accuracy of word analogy task in English. Mono refers to the monolingual English word embeddings.

The first baseline is Zhang et al. (2017a)[9], which perform adversarial training on transformation-based framework, and use "sharp drop" mechanism for model selection, k nearest neighbors (knn) to find translation pair. We fine-tune their model and then apply it on our datasets.

The other unsupervised baseline is Conneau et al. (2017)[10], they use a similar framework with Zhang et al. (2017a), and use CSLS for both model selection and translation pair finding, they show that CSLS significantly improves the accuracy on translation pair finding. We report the adversarial **without** refinement setting for fair comparison[11].

We also report the results of state-of-the-art seed-based models (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Dinu and Baroni, 2014; Artetxe et al., 2017; Smith et al., 2017) and Procrustes-CSLS (Conneau et al., 2017). These supervised baseline methods are trained with 5k seeds.

For word translation and sentence translation task, we report results of two methods for translation pair finding: Nearest Neighbors (NN) and CSLS.

4.4 Experimental Details

To ensure comparability, all the unsupervised models use 75k most common words in each language to feed the discriminator. At each training step, the word embeddings given to the discriminator are sampled uniformly. We consider the most frequency 200k word embeddings for evaluation.

Model Paramaters The encoder and decoder are linear layers as referred in subsection 3.2). Our discriminators are multilayer perceptrons with two hidden layers of size 2,500, and Leaky-ReLU activation functions. The input layer of discriminator is corrupted with a dropout rate of 0.1, the hidden layer of 0.1.

Training Details The autoencoders and discriminators are trained using stochastic gradient descent, with a learning rate of 0.1, and a mini-batch size of 32. A smoothing coefficient s = 0.1 is added to the discriminator predictions. We train discriminators more frequent than encoder-decoder with d_{step} = 5. We found λ_{recon} = λ_{cross} = λ_{adv} = 1 generally works well. α is tuned over the range [0,1] depending on language pairs.

4.5 Results

In this part, we first assess whether the induced corss-lingual word embeddings preserve the monolingual characteristics in Table 1. we present the main result on word translation comparing with our unsupervised baseline in Table 2 and comparison to supervised models in Table 3. The results on sentence translation retrieval task are given in Table 4.

Word Analogy To explore the monolingual characteristics of the induced cross-lingual word embedding, we apply it on word analogy task and present the result in Table 1. From Table 1, we show that by minimizing the self-reconstruction objective, our model effectively preserve the monolingual characteristics of the source embedding. When compared with others, we demonstrate that our model is comparable to other unsupervised models which try to preserve monolingual invariance by enforcing an orthogonality constraint on translation matrix.

Word translation We report the results on word translation task in Table 2. As it can be seen, results of the proposed model show consistent improvements over state-of-the-art models across most language pairs, whether take NN or CSLS as retrieval metrics. Moreover, we note that across language pairs which are substantially different such as English-Russian, English-Chinese, our model gets substantially better results, with up to 18% improvement in en-zh. These results match our argument that the interlingua-based method is more suitable for cross-lingual word representation learning.

Our method is particularly effective for low-resource languages given that the large parallel corpora are scarce in most language pairs. We

[9] We use their official released code http://www.thunlp.org/~zm/UBiLexAT/

[10] We use their official released code https://github.com/facebookresearch/MUSE

[11] Our model could also add refinement as a post-processing step.

Setting	Language pairs									
	en-es	es-en	en-fr	fr-en	en-ru	ru-en	en-zh	zh-en	en-eo	eo-en
Zhang17a	63.4	64.0	66.5	58.3	25.3	35.2	17.0	19.6	10.5	9.7
Conneau17-NN	69.9	71.3	70.4	61.9	29.1	41.5	18.5	22.3	13.5	12.1
Conneau17-CSLS	75.7	79.7	77.8	71.2	37.2	48.1	23.4	28.3	18.6	16.6
BiAAE-NN	73.3	73.7	75.0	73.9	35.5	49.6	35.0	28.5	15.9	14.3
BiAAE-CSLS	**78.5**	**80.5**	**79.9**	**79.1**	**40.6**	**55.3**	**37.0**	**33.3**	**21.0**	**18.6**

Table 2: Word translation retrieval accuracy in 5 language pairs for Wikipedia dataset. Best results are **bolded**.

Setting	en-it		it-en	
	P@1	P@5	P@1	P@5
Mikolov13†	33.8	48.3	24.9	41.0
Dinu15†	38.5	56.4	24.6	45.4
CCA†	36.1	52.7	31.0	49.9
Artetxe17†	39.7	54.7	33.8	52.4
Smith17†	43.1	60.7	38.0	**58.5**
Procrustes-CSLS*	**44.9**	68.1	**38.5**	57.2
Zhang17a	29.0	42.4	22.5	38.7
Conneau17-CSLS	40.2	54.7	35.2	53.5
BiAAE-CSLS	41.5	57.3	36.3	55.6

Table 3: Word translation retrieval accuracy (P@1, P@5) in English-Italian on the dataset of Dinu and Baroni (2014). Results marked with the symbol † are from Smith et al. (2017), symbol * is from Conneau et al. (2017). **Bold** indicates best results, underline indicates the best results among unsupervised models.

apply our model on English-Esperanto (en-eo) language pair, the top-1 accuracy of our model on en-eo is 21.0%, compared with 29.3% via the supervised method[12]. On eo-en, our unsupervised approach obtains 18.6%, compared with 25.3% in the supervised scenario. From a practical point, we report the top-5 accuracy, 41.5% and 39.7% for English-Esperanto and Esperanto-English respectively.

Since the previous results are based on comparable Wikipedia corpora, which is a comfortable setting. To address this, we carry out our evaluation on another more challenging dataset, and list the results of state-of-the-art seed-based supervised models for comparison (Table 3). As is shown to us, the proposed model achieves the best results among unsupervised models, which is similar to Table 2. The results of different datasets also confirm the robustness of the proposed method. In addition, when compared to supervised models, our model still get competitive results, with only 3.4% and 2.9% below the supervised approach in top-1 accuracy. This indicates that our approach provides a practical system for learning cross-lingual embeddings across close language pairs without any cross-lingual supervision.

Setting	en-it		it-en	
	P@1	P@5	P@1	P@5
Mikolov13†	10.8	18.7	12.0	22.1
Dinu15†	45.3	72.4	48.9	71.3
Smith17†	54.6	72.7	42.9	62.2
Procrustes-CSLS	**66.1**	**77.1**	**69.5**	**79.6**
Zhang17a	33.7	44.5	35.2	58.7
Conneau17-CSLS	42.5	57.6	47.0	62.1
BiAAE-CSLS	43.7	60.3	48.2	64.1

Table 4: Sentence translation retrieval accuracy (P@1, P@5) in English-Italian. Results marked with the symbol † are from Smith et al. (2017). **Bold** indicates best results. underline indicates the best results among unsupervised models.

Sentence translation Going from the word to the sentence level, we further evaluate the induced cross-lingual word embeddings on sentence translation retrieval task in Table 4. It can be easily seen that the proposed model generally have better performance than other two unsupervised methods. This suggests that for transferring semantic knowledge across languages via embeddings, our model is proved superior to previous unsupervised models.

5 Qualitative Analysis

In order to better understand the role of different objectives in the proposed model, we perform an ablation test, where we separately analyze the effect of the monolingual auto-encoding, the adversarial training, as well as the cross-lingual constraint. In practice, we separately remove each component from our framework and evaluate the monolingual characteristics and cross-lingual characteristics of the induced cross-lingual word embeddings. Table 5 show the obtained results.

From the 4th row, it's clearly that the

[12] Supervised results are from Conneau et al. (2017)

Setting	en-es		en-zh	
	WA	WT	WA	WT
Full system	80.2	78.5	79.5	37.0
- Mono.autoencoding	66.4	0.1	71.0	0.2
- Adv.training	78.3	0.0	77.8	0.0
- Cross.constraint	79.7	76.0	78.9	34.0

Table 5: Performance on Word Analogy (WA) and Word Translation (WT) tasks when individual component is removed (-).

monolingual auto-encoding plays a critical role in preserve the monolingual characteristics, the accuracy of word analogy drops severely when it is removed. Interestingly, we found our model even fails on cross-lingual tasks without this component. A possible explanation is that our model will stopping at a set of bad points without this constraint as a regularization. From the 5th row, we show that the adversarial training have an key influence on cross-lingual quality of the induced word embeddings. In the last row, we demonstrate that the cross-lingual constraint also have a positive influence on our model, with about 3% gain in performance in en-zh. This is consistent with our motivation proposed in 3.2.

In summary, every component of our model is indispensable to achieve better performances.

6 Related Works

6.1 Weakly Supervised And Unsupervised Models

Typically, a seed dictionary is designed to reduce the need of cross-lingual supervision. Frequent cognates and words, which are shared between two languages, were used to bootstrap translation pairs (Peirsman and Padó, 2010). Similarly, Smith et al. (2017) used identical character strings to form a parallel vocabulary. Artetxe et al. (2017) started their bootstrapping methods from a parallel vocabulary of aligned digits; while these methods still require cross-lingual evidence and not suitable to distant language pairs.

There are also other unsupervised attempts: Cao et al. (2016) explored a distribution-based method, which encourages word embeddings from different languages to lie in the shared semantic space by matching the mean and variance of the hidden states. Barone (2016) proposed an adversarial autoencoder-based model, which shares the common spirit with Zhang et al. (2017a,b); Conneau et al. (2017) and us. Although promising, the reported performance in both cases is poor in comparison to other methods. Another alternative way that does not rely on cross-lingual supervision is the decipherment approach (Dou et al., 2015). It views the source language as a cipher for the target language, and solves a statistical model that attempts to decipher the source language.

6.2 Adversarial Training

Generative adversarial networks have attracted a lot of attention since (Goodfellow et al., 2014), although it is originally designed for image generation, but adversarial training technique for matching distributions is generalizable to much more tasks, such as style transfer (Shen et al., 2017), representation learning (Xie et al., 2017), text generation (Hu et al., 2017).

Our approach is reminiscent of the Coupled GANs architecture (Liu and Tuzel, 2016) and its extensions proposed by Liu et al., while they implement the shared-latent space assumption using a weight sharing constraint to their encoders. In our proposals, we use independent autoencoders.

Another work similar to ours is Lample et al.'s work, which applies adversarial training technique to unsupervised neural machine translation. The difference is in following ways: (i) they use shared encoder-decoder while we not. (ii) they train one discriminator to distinguish the encoded vector while our model contains two discriminators which classify the cross-reconstructed vector and the source vector.

7 Conclusions and Future Work

In this work, we propose a novel Bilingual Adversarial AutoEncoder for unsupervised cross-lingual word representation learning. Based on results from our evaluation tasks, we show that by jointly transforming the source and target embeddings into a new common embedding space, the proposed model consistently outperforms previous unsupervised models on each cross-lingual task across various language pairs. In the future, we would apply our frame work to other downstream applications such as unsupervised machine translation and domain adaptation.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv preprint arXiv:1608.02996*.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *COLING*.

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

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

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113, Lisbon, Portugal. Association for Computational Linguistics.

Georgiana Dinu and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6568.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 836–845, Beijing, China. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

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

Stephan Gouws, Yoshua Bengio, and Gregory S. Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68, Baltimore, Maryland. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Controllable text generation. *CoRR*, abs/1703.00955.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *NIPS*.

Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *NIPS*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *VS@HLT-NAACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California. Association for Computational Linguistics.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NIPS*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard H. Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *NIPS*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.