A Bilingual Adversarial Autoencoder for Unsupervised Bilingual Lexicon Induction

Xuefeng Bai, Hailong Cao, Kehai Chen, and Tiejun Zhao

Abstract—**Unsupervised bilingual lexicon induction aims to generate bilingual lexicons without any cross-lingual signals. Successfully solving this problem would benefit many downstream tasks, such as unsupervised machine translation and transfer learning. In this work, we propose an unsupervised framework, named bilingual adversarial autoencoder, which automatically generates bilingual lexicon for a pair of languages from their monolingual word embeddings. In contrast to existing frameworks which learn a direct cross-lingual mapping of word embeddings from the source language to the target language, we train two autoencoders jointly to transform the source and the target monolingual word embeddings into a shared embedding space, where a word and its translation are close to each other. In this way, we capture the cross-lingual features of word embeddings from different languages and use them to induce bilingual lexicons. By conducting extensive experiments across eight language pairs, we demonstrate that the proposed method significantly outperforms the existing adversarial methods and even achieves best-published results across most language pairs.**

Index Terms—**Word embeddings, unsupervised word mapping, bilingual lexicon induction.**

I. Introduction

Unsupervised bilingual lexicon induction (**BiLI**) has attracted a lot of research interest. This interest stems from its natural property that the unsupervised bilingual lexicons could benefit thousands of low-resource languages which lack parallel corpus but have enough amount of monolingual ones, while at the same time providing inexpensive information for many downstream applications like unsupervised machine translation [1]–[5], information retrieval [6], text classification [7], [8], cross-lingual dependency parsing [9] and cross-lingual named entity recognition [10], [11].

In early works, it has been shown that word embeddings trained separately on monolingual corpora exhibit similar structures across languages [12], which suggests that different monolingual word embedding spaces could be connected with each other, allowing word features to transfer. Based on this finding, several recent works [13]–[18] have managed to induce bilingual lexicons from monolingual word embeddings without any cross-lingual signals. In spite of differences in detail, the common idea of these approaches is learning a cross-lingual word mapping which directly maps the source word embeddings to match the target word embeddings, then use the learned mapping to generate bilingual lexicons. However, such a direct cross-lingual word mapping might be suboptimal to generate high-quality bilingual lexicons. In reality, this approach relies heavily on the target word embedding space, which is pretrained on the target monolingual corpus and is not suitable for learning cross-lingual features.

In this paper, we propose to capture cross-lingual features of word embeddings from different languages in a third, latent space, then use the learned features to induce bilingual lexicons. To this end, we transform the source and the target word embeddings into a shared embedding space and constrain the transformed embeddings to be similar. The proposed model is implemented with a bilingual adversarial autoencoder (**BiAAE**) architecture, in which two autoencoders are trained jointly to transform the source and the target word embeddings into a shared embedding space. Meanwhile, the adversarial mechanism and distance-based training objectives are designed to guide the transformed word embeddings to be similar. To evaluate the effectiveness of our model, we conduct extensive experiments on two benchmark datasets across eight language pairs and compare our model with existing adversarial alternatives and other state-of-the-art supervised and unsupervised models. What's more, since previous adversarial systems are reported to fail to work on non-comparable datasets or distant language pairs [15], [19], we also perform experiments in these scenarios to verify the robustness of our model.

In summary, this paper makes the following contributions:

- We develop a novel adversarial framework (BiAAE) for unsupervised bilingual lexicon induction, which captures the cross-lingual features from two sets of word embeddings in a shared embedding space and use them to induce bilingual lexicons.
- Our model significantly outperforms existing adversarial alternatives and even achieves best-published results across most language pairs.
- Compared with previous adversarial approaches, the proposed model is more robust.

Manuscript received January 4, 2019; revised April 25, 2019 and June 11, 2019; accepted June 13, 2019. Date of publication July 3, 2019; date of current version July 22, 2019. This work was supported by the projects of National Natural Science Foundation of China under Grant 61572154 and in part by National Key Research and Development Program of China under Grant 2017YFB1002102. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hsin-min Wang. *(Corresponding author: Hailong Cao.)*

The authors are with the Machine Intelligence and Translation Laboratory, School of Computer Science of Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: xfbai.hk@gmail.com; caohailong2008@gmail.com; chenkehai@gmail.com; tjzhao@hit.edu.cn).

Digital Object Identifier 10.1109/TASLP.2019.2925973

II. Related Works

Recently, with the advance of continuous vector representation of words, a prevalent approach for bilingual lexicon induction is learning cross-lingual mappings of word embeddings [12]. We divide related works into the following three categories. Representative methods in each category are included in our comparative evaluation (Section IV-C).

A. Supervised Approaches

In early endeavors, most approaches for learning cross-lingual mappings are supervised. Mikolov *et al.* [12] first observed that monolingual word embeddings trained separately exhibit similar geometric properties across languages, which suggests that a linear mapping from the source word embedding space to the target word embedding space could be learned based on a seed dictionary and used for generating dictionaries and phrase tables. Faruqui and Dyer [20] used canonical correlation analysis **(CCA)** to learn two linear transformations for both sides that maximize the Pearson correlation coefficient. Motivated by a hypothetical inconsistency in [12], Xing *et al.* [21] incorporated length normalization in the training of word embeddings and maximized the cosine similarity, enforcing the orthogonality constraint to preserve the length normalization after mapping. Lazaridou *et al.* [22] proposed to learn a cross-lingual mapping which maximizes the margin between correct translations and rest of the candidates. Following previous work[12], Zhang *et al.* [23] used the same objective and constrained the mapping matrix to be orthogonal. Artetxe *et al.* [24] developed a framework which generalizes previous work and provides an efficient method to learn the optimal transformation. Nakashole [25] proposed to learn a transformation that is sensitive to the local neighborhood, which is particularly beneficial for distant languages. Doval *et al.* [26] showed that an additional transformation after the orthogonal alignment step could further improve the performance. Nevertheless, all these methods still require cross-lingual evidence.

B. Semi-Supervised Approaches

A related research line is to explore these frameworks for semi-supervised learning, where the seed lexicon is much smaller and used as a part of the bootstrapping process. Frequent cognates and words, which are shared between two languages, are used to bootstrap translation pairs [27]. Similarly, Smith *et al.* [28] used identical character strings to form a parallel vocabulary. Artetxe *et al.* [29] started their bootstrapping methods from a parallel vocabulary of aligned digits and obtained results that are comparable to those of supervised methods when starting with only 25 word pairs. However, these methods make strong assumptions on the writing systems of languages and are not suitable for distant languages (e.g., English-Arabic).

C. Unsupervised Approaches

Recently, several fully unsupervised approaches have been proposed for learning cross-lingual word mappings. A typical research line is based on adversarial training, commonly known as Generative Adversarial Networks **(GANs)** [30]. Barone [31] proposed an adversarial autoencoder-based model, where an encoder maps the source language word embeddings into the target language, a decoder reconstructs the source language word embeddings from the mapped embeddings, and a discriminator distinguishes between the mapped word embeddings and the real target language word embeddings. Although promising, the reported performance is not satisfying. Zhang *et al.* [16] incorporated the cross-lingual word mapping architecture with adversarial training and achieved very promising results. Zhang *et al.* [17] adopted the earth mover's distance for training, optimized through a wasserstein generative adversarial network followed by an alternating optimization procedure. Conneau *et al.* [18] took the adversarial framework as initialization and combined it with a self-learning process. They reported encouraging results on a large dataset. Based on the previous framework [18], Chen *et al.* [32] proposed to induce the bilingual lexicon in multilingual scenarios. A recent work similar to us is the method of Dou *et al.* [33]. They made a latent space assumption and their work combined variational autoencoders **(VAE)** [34] and GANs. In comparison with this work, we propose a different framework and introduce new training objectives.

On the other hand, non-adversarial approaches have also been proposed for unsupervised BiLI. For instance, Mukherjee *et al.* [35] proposed a statistical dependency-based approach which searches for the cross-lingual word pairing that maximizes statistical dependency in terms of squared loss mutual information (**SMI**). Artetxe *et al.* [15] explored the similarity matrix to learn an unsupervised initial weak cross-lingual word mapping and combined it with a robust self-learning approach. Similarly, Aldarmaki *et al.* [13] proposed to use the similarity of the k-nearest-neighbor graph to learn an unsupervised word mapping and refine it subsequently. Hoshen and Wolf [14] used principal component analysis **(PCA)** to learn an initial alignment and then iteratively refine the alignment. Xu *et al.* [36] adopted the Sinkhorn distance and incorporated back-translation to transfer a word embedding space to another. Alvarez-Melis and Jaakkola [37] cast the problem as an optimal transport **(OT)** problem directly and utilized Gromov-Wasserstein distance to measure the similarities between pairs of words across languages. Our framework learns the word mapping in an adversarial way and could be further combined with the helpful techniques used in non-adversarial approaches.

III. The Bilingual Adversarial Autoencoder Model

We start this section by introducing motivations and intuitions behind our approach. Next, we formally present the model and the unsupervised criterion for model selection.

Notation: Throughout this work, we denote the source and the target monolingual word embeddings as $X_1 \subset \mathbb{R}^{d_{X_1}}$ and $X_2 \subset \mathbb{R}^{d_{X_2}}$, respectively, where d_{X_1} and d_{X_2} are dimension of word embeddings. For simplicity of representation, we assume that $d_{X_1} = d_{X_2}$. Let X_{1Z} and X_{2Z} denote the transformed word embeddings. The encoder of the source and the target language are denoted as Enc_1, Enc_2, respectively. Similarly, Dec_1, Dec_2

denote the decoder of the source and the target language, respectively.

A. Motivation and Intuitions

In this paper, we focus on capturing the shared features of word embeddings in a latent space without any cross-lingual supervisions. To this end, we proposed to: (i) transform the source and the target word embeddings into a shared embedding space by autoencoders, (ii) encourage the transformed word embeddings be similar.

Our intuitions are implemented as:

i) *Monolingual Autoencoding:* Two auto-encoders are used to transform monolingual word embeddings into a shared embeddings space while preserving original information as much as possible. This is implemented by standard autoencoders [38]. More specifically, the Enc_1 (or Enc_2) encodes monolingual word embeddings X_1 (or X_2) into X_{1Z} (or X_{2Z}), while the Dec_1 (or Dec_2) learns to reconstruct word embeddings.

ii) *Adversarial Mechanism:* Since we do not have access to the word correspondence across languages, we propose to match the transformed embeddings X_{1Z} and X_{2Z} at the distribution level. Our assumption is that: if 1) the distribution of the source word embeddings X_1 can be "cross-reconstructed" from the transformed word embeddings of the target language X_{2Z}; 2) the distribution of the target word embeddings X_2 can be "cross-reconstructed" from the transformed word embeddings of the source language X_{1Z}, then the distributions of transformed word embeddings X_{1Z} and X_{2Z} are considered to be matched. To satisfy the above two constraints, we use adversarial loss as a discrepancy criterion. Specifically, a discriminator D_1 for the source language is introduced to classify between the cross-reconstructed word embeddings (denoted as $\tilde{X}_{2Z\to 1}$) and the source word embeddings X_1. Similarly, a discriminator D_2 for the target language is used to classify between the cross-reconstructed word embeddings (denoted as $\tilde{X}_{1Z\to 2}$) and the target word embeddings X_2.

iii) *Cross-lingual Constraint:* Since the discriminator D_1 and D_2 could inevitably "overfit", which is harmful to the alignment of X_{1Z} and X_{2Z}, we introduce another distance-based learning objective ℓ_{cross} to alleviate this problem.

The total architecture of our model is illustrated in Fig. 1.

B. Learning Objectives

Following the intuitions mentioned in the previous section, this section will introduce the training objectives of the proposed model in detail.

1) Monolingual Autoencoding: Take the source language as an example, the encoder Enc_1 takes the source word embeddings as input and generates latent (transformed) word embeddings X_{1Z}. The decoder Dec_1 makes use of the transformed word embeddings X_{1Z} and produces the reconstructed word embeddings $\tilde{X}_{1Z\to 1}$. Then, at each iteration, the encoder and decoder are trained to minimize a loss function which measures the discrepancy between original and the reconstructed word embeddings:

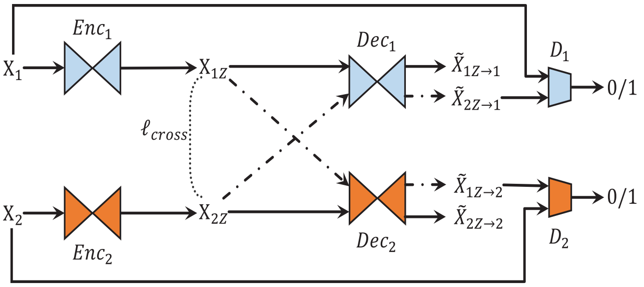

Fig. 1. The proposed BiAAE framework. Two encoders Enc_1 and Enc_2 are used to transform the source word embeddings X_1 and the target word embeddings X_2 into a shared embedding space. Two decoders Dec_1 and Dec_2 are designed to reconstruct (solid arrow) and cross-reconstruct (dotted arrow) word embeddings. Note that X_{1Z} and X_{2Z} are transformed word embeddings, $\tilde{X}_{1Z\to 1}$ and $\tilde{X}_{2Z\to 2}$ are reconstructed word embeddings; $\tilde{X}_{2Z\to 1}$ and $\tilde{X}_{1Z\to 2}$ are cross-reconstructed word embeddings. D_1 and D_2 are discriminators.

$$\begin{aligned}\ell_{mono}(X_1) &= \mathbb{E}_{x_1\sim X_1}[\triangle(x_1, \tilde{x}_{1Z\to 1})]\\ &= \mathbb{E}_{x_1\sim X_1}[\triangle(x_1, Dec_1(Enc_1(x_1)))], \qquad (1)\end{aligned}$$

where $x_1 \sim X_1$ denotes that x_1 is sampled from X_1, $\tilde{x}_{1Z\to 1}$ is the reconstructed word embedding of x_1, and $\triangle$ denotes the discrepancy criterion, which is set as the average cosine similarity in our model.

Similarly, for the target language, the following loss function is minimized:

$$\begin{aligned}\ell_{mono}(X_2) &= \mathbb{E}_{x_2\sim X_2}[\triangle(x_2, \tilde{x}_{2Z\to 2})]\\ &= \mathbb{E}_{x_2\sim X_2}[\triangle(x_2, Dec_2(Enc_2(x_2)))], \qquad (2)\end{aligned}$$

2) Adversarial Mechanism: Recent works have shown that adversarial training could be utilized to match two distributions without any supervision [3], [16]–[18]. In this work, we adopt this idea to guide the transformed word embeddings X_{1Z} and X_{2Z} to be similar. Specifically, we first "cross-reconstruct" the target word embddings from the source transformed word embeddings X_{1Z} by Dec_2, and vise versa:

$$\begin{aligned}\tilde{X}_{1Z\to 2} &= Dec_2(X_{1Z}) = Dec_2(Enc_1(X_1)),\\ \tilde{X}_{2Z\to 1} &= Dec_1(X_{2Z}) = Dec_1(Enc_2(X_2)). \qquad (3)\end{aligned}$$

Then we utilize adversarial training to match the distributions of $\tilde{X}_{1Z\to 2}$ and X_2, similarly $\tilde{X}_{2Z\to 1}$ and X_1. The adversarial training is formalized as the following minimax game:

1) The neural network-based discriminators D_1 and D_2 are trained to classify the original word embeddings and the cross-reconstructed word embeddings by maximizing the following objective:

$$\begin{aligned}\ell_{D_1}(X_1, X_2) &= \mathbb{E}_{x_1\sim X_1}[\log D_1(x_1)]\\ &\quad + \mathbb{E}_{x_2\sim X_2}[\log(1 - D_1(Dec_1(Enc_2(x_2))))], \qquad (4)\end{aligned}$$

$$\begin{aligned}\ell_{D_2}(X_1, X_2) &= \mathbb{E}_{x_2\sim X_2}[\log D_2(x_2)]\\ &\quad + \mathbb{E}_{x_1\sim X_1}[\log(1 - D_2(Dec_2(Enc_1(x_1))))]. \qquad (5)\end{aligned}$$

2) The two autoencoders are trained to confuse the discriminators by minimizing the following objective:

$$\ell_{adv}(X_2) = \mathbb{E}_{x_2 \sim X_2}[\log(1 - D_1(Dec_1(Enc_2(x_2))))], \tag{6}$$

$$\ell_{adv}(X_1) = \mathbb{E}_{x_1 \sim X_1}[\log(1 - D_2(Dec_2(Enc_1(x_1))))]. \tag{7}$$

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

3) Cross-Lingual Constraint: Note that there still exists "pseudo" word embedding pairs (x_{1Z}, x_{2Z}) that could confuse the discriminators since the discriminator D_1 and D_2 could "overfit" (such as focusing on a combination of local differences between the distributions), which is a common problem of GANs.[1] To tackle this issue, the following distance-based objective is proposed to encourage (x_{1Z}, x_{2Z}) be similar as much as possible:

$$\ell_{cross}(X_1, X_2) = \mathbb{E}_{x_1 \sim X_1, x_2 \sim X_2}[\triangle(Enc_1(x_1), Enc_2(x_2))], \tag{8}$$

where x_1, x_2 is sampled from X_1 and X_2 separately, $\triangle$ denotes the criterion function, we use average cosine distance in practice.

Total objective function: The total objective function of autoencoders is:

$$\ell_{total} = \lambda_{modo}(\ell_{mono}(X_1) + \ell_{mono}(X_2)) + \lambda_{adv}[\ell_{adv}(X_1) + \ell_{adv}(X_2)] + \lambda_{cross}\ell_{cross}(X_1, X_2), \tag{9}$$

where λ_{mono}, λ_{cross} and λ_{adv} are weighting hyper-parameters for ℓ_{mono}, ℓ_{cross} and ℓ_{adv}, respectively. At each iteration, we optimize the autoencoder loss (Equation (9)) and discriminator loss (Equation (4), (5)) alternately.

C. Model Selection

Since the adversarial models are difficult to converge, we wish to have a criterion correlated with the quality of induced dictionary to select the best model. However, we do not have access to parallel data to judge how well our model works, not even at validation time. In previous works, Zhang *et al.* [16] proposed to use "sharp drops" of generator loss to select a model, while Conneau *et al.* [18] introduced an unsupervised criterion which quantifies the closeness of produced word embeddings.

In this work, we adopt the criterion proposed by Conneau *et al.* [18] to perform model selection since it works well on large scale dataset. We consider the 10k most frequent source words and use cross-domain similarity local scaling (**CSLS**) to select a translation for each word, and compute the average cosine similarity between each word and its translation. Finally, this average is used as a validation metric. Given two word embeddings x_{1Z} and x_{2Z}, the CSLS similarity score is then calculated as follows:

$$\text{CSLS}(x_{1Z}, x_{2Z}) = 2cos(x_{1Z}, x_{2Z}) - \tau(x_{1Z}) - \tau(x_{2Z}) \tag{10}$$

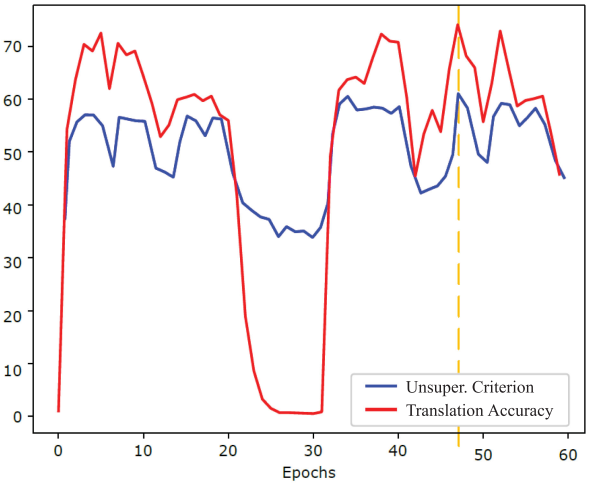

Fig. 2. Unsupervised criterion and actual BiLI accuracy in our experiment, we select the point where CSLS reaches to a maximum value (e.g., epoch 47).

where $\tau(x_{1Z})$ and $\tau(x_{2Z})$ are the average cosine similarity of x_{1Z} (or x_{2Z}) to its k nearest neighbors,[2] respectively.

We show that this criterion is well correlated with the performance in Fig. 2.

IV. Experiments

In this section, we first introduce the detailed experimental setup, including the datasets, evaluation metrics, baselines, and implementation details. Then we empirically evaluate the effectiveness of our unsupervised approach on standard benchmark datasets and compare it with existing adversarial models (Section V-A). We also test the proposed model on more realistic scenarios (Section V-B). At last, we combine our model with some helpful techniques and compare it with other state-of-the-art unsupervised and supervised approaches (Section V-C).

A. Datasets

We conduct our experiments on 2 benchmark datasets: i) The Multilingual Unsupervised and Supervised Embeddings **(MUSE)** dataset released by Conneau *et al.* [18]. This dataset consists of monolingual word embeddings of dimension 300 trained with fastText[3] on Wikipedia monolingual corpora[4] and gold dictionaries for 110 language pairs. Here, we focus on the language pairs for which they report results: English (EN) from/to Spanish (ES), French (FR), German (DE), Russian (RU), Chinese (ZH), and Esperanto (EO). ii) The **Vecmap** dataset from Dinu and Baroni [39] and the extensions of Artetxe *et al.* [29]. This dataset consists of gold dictionaries and 300-dimensional CBOW[5] embeddings trained on WacKy crawling corpora (English, Italian, German), Common Crawl (Finish) and WMT News Crawl (Spanish). We report the results

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

[2] We set $k = 10$.

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

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

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

on the following four pairs of languages: EN from Spanish (ES), Italian (IT), German (DE), and Finnish (FI).

B. Evaluation Metrics

We first generate bilingual lexicons through learned word embeddings (X_{1Z} and X_{2Z} in Fig. 1), then measures the accuracy of the induced dictionary in comparison to a gold standard. 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 similarity of induced word vectors. We use the high-quality gold dictionaries mentioned previously (Section IV-A) for evaluation and adopt standard evaluation procedure. For each language pair, we consider 1,500 query source and 200k target words. Top-1 accuracy is reported in this task. We utilize two metrics for retrieving the top-1 nearest neighbors: Nearest Neighbors (NN) and CSLS (Section III-C).

C. Baselines

We evaluate our method in comparison with the following state-of-the-art models:

- **Adv-Z**,[6] an adversarial method proposed by Zhang *et al.* [16]. It combines the cross-lingual word mapping framework with adversarial training and uses "sharp drop" mechanism for model selection.
- **Adv-C**,[7] an adversarial approach proposed by Conneau *et al.* [18]. It shares a similar framework with Zhang *et al.* [16] and utilize CSLS for both model selection and word translation retrieval. It can further be combined with Procrustes Analysis to refine the learned mapping.
- **Adv-latent**, an adversarial model proposed by Dou *et al.* [33]. It combines VAE and GANs to learn latent variables that could capture the semantic meaning of words. We reproduce their model and evaluate it on all tested language pairs.
- **G-W**, a non-adversarial approach proposed by Melis and Jaakkola [37]. It optimizes the Gromov-Wasserstein distance of words across languages. We use the results reported in their paper.
- **Non-adv-H**, a non-adversarial approach proposed by Hoshen and Wolf [14]. It uses PCA for initial alignment and an iterative learning method for refinement. The accuracies are taken from their paper.
- **Non-adv-A**,[8] a robust non-adversarial approach proposed by Artetxe *et al.* [15]. It includes an unsupervised initial weak mapping and a robust self-learning refinement.
- We also report the results of some seed-based supervised models: Transformation Matrix **(TM)** [12], Orthogonal Transformations with inverted softmax function **(OT-ISF)** [28], **Advanced-Mapping** [24] and **Procrustes-CSLS** [18]. All supervised baseline models are trained with 5k seed word translation pairs.

D. Implementation Details

To ensure comparability, all the adversarial models use 75k most frequent words in each language to feed the discriminator. At each training step, the word embeddings given to the discriminator are sampled uniformly.[9] We consider the most frequent 200k word embeddings for evaluation. It should be noted that all results reported in the paper are an average of 10 runs. Our implementation is released as an open source project at https://github.com/muyeby/BiAAE.

Model Parameters: The encoder and decoder are single linear layers. The discriminators are multilayer perceptrons with 2 hidden layers of size 2,500, and a Leaky-ReLU activation function. The input layer of discriminator is corrupted with a dropout rate of 0.1. We found $\lambda_{mono} = \lambda_{cross} = \lambda_{adv} = 1$ generally works well.

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 frequently (5 times) than autoencoders.

V. Results

A. Main Results

The main results evaluated on MUSE dataset are given in Table I. We compare our model with state-of-the-art adversarial systems (Adv-Z, Adv-C, and Adv-latent). Clearly, the proposed model significantly outperforms other adversarial methods on 11 of 12 language pairs when we take NN as the retrieval metrics. The similar trend is also observed when using CSLS. More importantly, we note that across language pairs like English-French (EN-FR), English-Russian (EN-RU) and English-Chinese (EN-ZH), our model gets substantially better results than Adv-C, with up to 8% improvement in EN-ZH. The reason why the proposed model gets comparable results with Adv-C on ES-EN, but much better results on the opposite directions is that our method tends to find a shared space which is suitable for both forward and backward BiLI tasks, instead of focusing on a single direction. Besides, our method obtains significantly better results than Adv-latent which shares a similar idea with us, especially on EN-{FR, DE, RU, ZH}. We further discuss this in Section VI. All these results suggest that the proposed approach can generate higher quality dictionaries than previous adversarial frameworks.

Table II shows the average cosine similarity of top-5000 frequency English words and their translations in other languages. Compared with Adv-C, the proposed model can make translated words closer. That is to say, word embeddings can be aligned easier in a shared latent space than original monolingual space.

Table III presents several case studies. In the first example, both methods find the correct translations. In the following three examples, our approach successfully induces the correct translation while Adv-C fails, although these words have significantly different meanings. These examples indicate that

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

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

[8] https://github.com/artetxem/Vecmap

[9] We also tried sampling by word frequency, without observing any significant improvement.

TABLE I
BILINGUAL LEXICON INDUCTION ACCURACY (TOP-1) ACROSS 6 LANGUAGE PAIRS ON **MUSE**. †RESULTS AS REPORTED IN THE ORIGINAL PAPER. BEST RESULTS ARE **BOLDED**

	EN-ES		EN-FR		EN-DE		EN-RU		EN-ZH		EN-EO	
Model	→	←	→	←	→	←	→	←	→	←	→	←
Adv-Z	63.4	64.0	66.5	58.3	40.3	41.4	25.3	35.2	17.0	19.6	10.5	09.7
Adv-C-NN	69.9	71.3	70.4	61.9	61.5	59.6	29.1	41.5	18.5	22.3	13.5	12.1
Adv-C-CSLS	75.7	**79.7**	77.8	71.2	69.2	66.4	37.2	48.1	23.4	28.3	18.6	16.6
Adv-latent-NN	71.8†	71.0	71.1	67.5	63.7	60.6	32.8†	41.4	22.9†	23.3	14.0	12.5
Adv-latent-CSLS	76.6†	77.1	76.3	74.1	68.3	67.3	39.3†	44.9	26.0†	28.7	19.2	18.0
Ours-NN	73.8	71.7	75.3	71.4	63.8	65.8	37.3	46.5	26.4	28.5	15.0	14.3
Ours-CSLS	**77.3**	79.4	**80.0**	**76.5**	**69.8**	**68.7**	**40.7**	**48.9**	**30.5**	**31.3**	**20.2**	**18.6**

TABLE II
AVERAGE COSINE SIMILARITY OF TRANSLATED WORDS ON **MUSE**. RESULTS ARE BASED ON CSLS

	Language Pairs					
Model	EN-ES	EN-FR	EN-DE	EN-RU	EN-ZH	EN-EO
Adv-C	0.71	0.73	0.69	0.63	0.62	0.51
Ours	**0.73**	**0.74**	**0.70**	**0.68**	**0.69**	**0.60**

TABLE III
WORD TRANSLATION EXAMPLES FOR ENGLISH-ITALIAN ON **MUSE**. RESULTS ARE BASED ON CSLS

Query	Adv-C	Ours	Gold
three	tre	tre	tre
neck	toracica	collo	collo
door	finestrino	portiera	portiera
room	bagno	camera	camera
before	dopo	dopo	prima

TABLE IV
ACCURACY (TOP-1) OF UNSUPERVISED APPROACHES ON **VECMAP** [29], [39]

	Language Pairs			
Model	EN-DE	EN-IT	EN-ES	EN-FI
Adv-Z	00.00	00.00	00.01	00.01
Adv-C-NN	33.62	08.57	15.31	00.01
Adv-C-CSLS	37.15	13.55	18.23	00.38
Ours-NN	36.20	33.20	22.20	17.42
Ours-CSLS	**40.40**	**38.90**	**22.82**	**21.49**

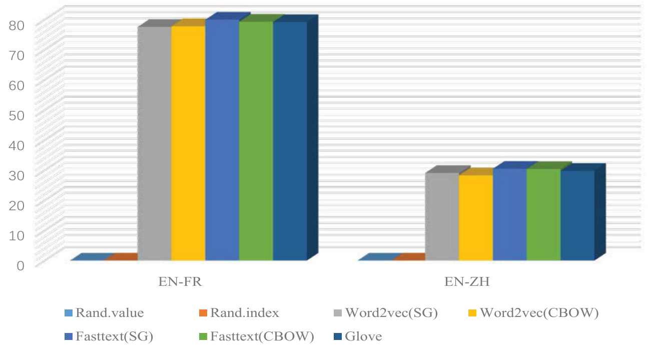

Fig. 3. Bilingual Lexicon Induction Accuracy on MUSE with regard to Word2vec (skip-gram), Word2vec (continuous bag-of-words), Fasttext (skip-gram), Fasttext (continuous bag-of-words) or Glove [40] word embeddings.

the direct mapping-based method is sub-optimal for aligning monolingual word embeddings and our approach is better for that. In the last example, both methods fail to generate the correct translation but find an antonym. A possible explanation for this phenomenon is that antonyms might have similar word embeddings as they always have similar contexts.

B. Robustness Test

Although previous adversarial adventures have reported promising results on BiLI, recent works have shown that these models have important limitations [15], [19]. When tested on more realistic scenarios, the performance of such models drops dramatically [15]. To present an extensive evaluation of our approach, we i) apply the proposed model on more challenging[10] datasets released by Dinu and Baroni [39] and Artetxe *et al.* [29]. ii) feed our method with word embeddings pre-trained by different algorithms.

Impact of the corpus: We first test whether our model is sensitive to training corpus. As shown in Table IV, Adv-Z fails in strictly monolingual conditions, which is consistent with the negative results reported by the authors themselves for similar conditions. Although Adv-C can report positive results[11] on EN-DE, EN-IT, and EN-ES, the performance is far from that reported in MUSE dataset. In distant language pair En-FI, Adv-C nearly fails (only 1.62% in the best run). This phenomenon is consistent with results in recent works [19]. In contrast, our method is much more robust. It can still achieve very promising results on such challenging datasets, no matter the language pairs are close-related (EN-DE, IT, ES) or distant (EN-FI).

Impact of the pre-training methods: We then investigate whether the performance of our model is influenced by pre-training algorithms. We consider following pre-training algorithms: i) Word2vec (SG/CBOW), ii) FastText (SG/CBOW), iii) Glove [40]. For the convenience of comparison, we add two types of "random" word embeddings as baselines: i) we assign each word with random embeddings instead of pre-trained embeddings (denoted as Rand.value). ii) we assign each word with embedding of other words (denoted as Rand.index). As shown in Fig. 3, the Rand.value embeddings almost fail to work on both EN-FR and EN-ZH, with 0.0% on EN-FR and 0.1% on EN-ZH. The Rand.index embeddings perform similarly, with 0.3% on EN-FR and 0.3% on EN-ZH. These results indicate

[10] It should be noted that all sentences are collected from strictly monolingual corpora, and the distant language pair English-Finnish is more challenging.

[11] Note that across these language pairs, the Adv-C could only succeed about 6 times in 10 runs.

TABLE V
BILINGUAL LEXICON INDUCTION ACCURACY (TOP-1) OF THE PROPOSED METHOD IN COMPARISON WITH PREVIOUS WORK ON **MUSE**. BEST RESULTS ARE **BOLDED**. BEST RESULTS AMONG ADVERSARIAL METHODS ARE UNDERLINED. "REFINE" REFERS TO THE SELF-LEARNING PROCEDURE. †RESULTS AS REPORTED IN THE ORIGINAL PAPER. *FAILED TO CONVERGE

	EN-ES		EN-FR		EN-DE		EN-RU		EN-ZH		EN-EO	
Model	→	←	→	←	→	←	→	←	→	←	→	←
Supervised methods (5k dict.)												
TM	72.5	74.0	70.9	71.3	61.2	61.9	44.3	50.3	13.3	29.9	16.3	16.4
Advanced-Mapping	79.8	82.1	80.5	81.3	72.7	71.7	50.1	65.1	38.3	28.1	28.1	**31.7**
Procrustes-CSLS	81.4	82.9	81.1	82.4	73.5	72.4	**51.7**	63.7	42.7	36.7	29.3	25.3
Unsupervised non-adversarial methods												
G-W ($\lambda = 10^{-5}$)	81.7	80.4	81.3	78.9	71.9	72.8	45.1	43.7	-	-	-	-
Non-adv-H-NN	75.9	76.0	74.8	75.0	66.9	67.1	36.8	48.4	*	*	-	-
Non-adv-H-CSLS	81.1	82.1	81.5	81.3	73.7	72.7	44.4	55.6	*	*	-	-
Non-adv-A-NN	80.5	81.5	80.5	80.1	73.3	72.3	44.8	62.1	33.5	33.5	23.2	25.2
Non-adv-A-CSLS	**82.3**	**84.8**	82.3	83.6	**75.2**	**74.1**	49.1	65.5	37.5	**37.8**	**30.7**	28.6
Unsupervised adversarial methods												
Adv-C-NN-refine	79.1	78.1	78.1	78.2	71.3	69.6	37.3	54.3	30.9	21.9	20.7	20.6
Adv-C-CSLS-refine	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4	28.2	25.6
Adv-latent-NN-refine	79.1†	79.3	78.7	77.4	70.9	70.7	42.7†	53.9	32.5†	24.5	22.8	21.0
Adv-latent-CSLS-refine	82.1†	83.9	81.9	82.5	74.1	73.5	48.7†	60.1	33.3†	32.1	29.0	25.9
Ours-NN-refine	80.6	81.8	80.8	80.2	73.5	72.5	44.5	62.8	36.3	33.2	26.5	22.9
Ours-CSLS-refine	**82.3**	84.3	**82.5**	**83.7**	**75.2**	**74.1**	49.0	**65.8**	**43.4**	36.7	30.2	26.6

that pre-trained embeddings are indispensable for our model. Besides, the performance of Word2vec (SG) embeddings and Word2vec (CBOW) embeddings are comparable in both EN-FR and EN-ZH. The FastText (SG/CBOW) embeddings obtain two percent more accuracy compared to Word2vec (SG/CBOW) embeddings trained on the same corpus. This improvement is likely because that FastText embeddings incorporate more syntactic information about the words. Moreover, the performance of Glove embeddings is slightly better than Word2vec (SG/CBOW) embeddings, because the former contains the global information of words.

C. *Effect of Self-Training*

Inspired by Artetxe *et al.* [29], we further combine our system with self-learning[12] which iteratively refines the mapping. In practice, we induce a synthetic bilingual lexicon using the learned adversarial model. Subsequently, we use this generated dictionary as anchors to improve the mapping iteratively. To have a systematic comparison, we carry out experiments on both *easier* (MUSE) and *harder* (Vecmap) datasets. The results are presented in Table V and Table VI, respectively.

We first compare our approach with state-of-the-art adversarial systems (last few lines in Table V and Table VI). It is clear that the proposed method shows consistent improvement than Adv-Z, Adv-C, and Adv-latent. Moreover, while the performance of Adv-Z and Adv-C drops dramatically in Vecmap or distant languages (EN-FI), our method could consistently report high accuracies. When compared with other non-adversarial systems, the results show that: (1) our model could perform competitively and even superiorly compared with state-of-the-art non-adversarial approaches, whether on MUSE or Vecmap. (2) the proposed model even achieves best-published results among unsupervised approaches on EN-FR (MUSE), RU-EN (MUSE), EN-ZH (MUSE), EN-FI (Vecmap) and EN-ES (Vecmap).

We also report the results of several state-of-the-art supervised models trained with 5k seeds. In Table V, one can see that the proposed model even surpasses previous strong supervised systems on EN-{ES, FR, DE}, and gets close results on EN-RU. Table VI shows similar trends. This is especially impressive for the low-resource languages and domains.

TABLE VI
BILINGUAL LEXICON INDUCTION ACCURACY (TOP-1) OF THE PROPOSED METHOD IN COMPARISON WITH PREVIOUS WORK ON **VECMAP**. BEST RESULTS ARE **BOLDED**. BEST RESULTS AMONG ADVERSARIAL METHODS ARE UNDERLINED. "REFINE" REFERS TO THE SELF-LEARNING STEPS

Settings	EN-IT	EN-DE	EN-FI	EN-ES
Supervised methods (5k dict.)				
TM	34.93	35.00	25.91	27.73
OT-ISF	43.13	43.33	29.42	35.13
Advanced-Mapping	45.27	44.13	32.94	36.60
Procrustes-CSLS	45.67	47.27	32.38	37.00
Unsupervised non-adversarial methods				
G-W-NORMALIZE	**49.21**	46.50	18.30	37.60
Non-adv-A-NN	44.12	45.43	29.90	33.23
Non-adv-A-CSLS	48.01	**48.22**	32.70	37.47
Unsupervised adversarial methods				
Adv-Z	00.00	00.00	00.07	00.07
Adv-C-NN-refine	40.56	41.37	00.57	32.48
Adv-C-CSLS-refine	45.40	46.83	01.62	37.08
Ours-NN-refine	43.70	44.07	30.30	34.27
Ours-CSLS-refine	47.50	47.87	**33.08**	**38.27**

VI. QUALITATIVE ANALYSIS

A. *Ablation Test*

In order to better understand the role of different objectives in our model, we perform an ablation test, where we separately analyze the effect of the monolingual auto-encoding, the adversarial mechanism, as well as the cross-lingual constraint. In practice, we separately remove each component from our framework and evaluate the model on word analogy [41] and bilingual lexicon induction. The obtained results on MUSE dataset are presented in Table VII.

From the 4th row, it could be observed that the monolingual auto-encoding plays a critical role in preserving the monolingual characteristics—the accuracy of word analogy drops severely

[12] We adopt the robust self-learning techniques proposed by Artetxe *et al.* [15]

TABLE VII
PERFORMANCE ON WORD ANALOGY (WA) AND BILINGUAL LEXICON INDUCTION (BILI) WHEN INDIVIDUAL COMPONENT IS REMOVED (-) OR ADDED (+). MONO. BASELINE REFERS TO THE ORIGINAL MONOLINGUAL WORD EMBEDDINGS. ALL BILI RESULTS ARE BASED ON CSLS

	EN-ES		EN-FR		EN-DE		EN-RU		EN-ZH		EN-EO	
Setting	WA	BiLI	WA	BiLI	WA	BiLI	WA	BiLI	WA	BiLI	WA	BiLI
Mono. baseline	**80.3**	-	**80.3**	-	**80.3**	-	**80.3**	-	**80.3**	-	**80.3**	-
Full system (without refine)	80.2	**77.3**	79.9	**80.0**	80.0	69.8	80.1	40.7	79.5	**30.5**	79.8	20.2
- Mono.autoencoding	66.4	00.1	68.3	00.2	67.5	00.2	65.3	00.2	71.0	00.2	65.7	00.1
- Adv.mechanism	78.3	00.0	78.3	00.0	78.5	00.1	76.9	00.0	77.8	00.0	78.0	00.0
- Cross.constraint	79.7	76.8	79.9	78.2	79.2	69.0	79.3	39.8	78.9	28.4	79.5	19.5
+ Ortho.constraint	80.2	76.2	80.1	78.5	79.9	67.9	80.0	38.3	79.7	28.8	80.1	19.1

when it is removed. This phenomenon is in line with our intuitions. Additionally, the proposed model almost fails in BiLI without this component. We attribute this to the positive influence of this component on avoiding the local optima.

As for the adversarial constraint, we observe that it has a crucial influence on unsupervised BiLI, only in less than 1/1000 cases does a nearest neighbor search return a correct translation when it is removed.

Moreover, we demonstrate that the cross-lingual constraint also has a positive influence on our model, with 1.8 points gain in EN-FR, 0.9 points gain in EN-RU, 2.1 points gain in EN-ZH. This is consistent with our motivation mentioned in the previous section.

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

B. Discussion: Why it Works

The latent space model: The key idea of the proposed model is to introduce a third, latent space rather than the source (target) space to capture the cross-lingual features of word embeddings. This stems from the fact that a direct cross-lingual word mapping is not an optimal solution to capture the shared features. Interestingly, some recent works toward supervised cross-lingual word representation learning came to similar conclusions. For instance, Doval *et al.* [26] showed that the current framework could be further improved by meeting the mapped word embeddings and the target word embeddings in the middle. Kementchedjhieva *et al.* [42] showed that mapping the languages onto a third space rather than directly onto each other makes it easier to learn an alignment.

The form of the mapping: The other improvement in our model is that we adopt the linear mapping with autoencoder constraint rather than orthogonal mapping. Existing unsupervised frameworks [15], [16], [18], [24], [33], [42] use orthogonal mapping to align word embedding spaces of different languages, under the assumption that the internal structures of different monolingual word embeddings spaces are approximately isomorphic [31]. However, recent researches [19], [42] have shown that word vector spaces are often relatively far from being isomorphic, and approximate isomorphism is not transitive. For this reason, we propose to use a more expressive mapping with autoencoder constraint. In fact, we test both orthogonal mapping and non-orthogonal mapping, the model based on non-orthogonal (linear) mapping can outperform model parameterized by orthogonal mapping with the same number of parameters (see the last line in Table VII). Such results are in line with several recent works which argue that given the important differences (e.g., vocabulary, grammar, written form, or syntax) between different languages, more effective mapping instead of orthogonal mapping should be utilized [43]–[45]. We further test nonlinear mappings parameterized by neural networks, although obtain promising results sometimes, we found the overall performance is very unstable. This phenomenon is likely to be solved from advances in techniques that further stabilize adversarial training.

Comparing with Dou et al. [33]: Our method is similar to the work of Dou *et al.* [33]. The proposed approach has two advantages in comparison with Dou *et al.* [33]: Firstly, the proposed model has two discriminators, which pushes the autoencoders to generate more indistinguishable latent word embeddings than Dou *et al.* [33]. By comparing the results in Table VII and Table I, we observe that the adversarial part (without cross-lingual constraint) of the proposed model can give better reuslts than model of Dou *et al.* [33]. Secondly, we additionally introduce the cross-lingual constraint, which is useful to solve the problem caused by “overfitting” of GANs. We have verified that this constraint can further improve performance of the proposed model (See Section VI-A).

VII. CONCLUSIONS AND FUTURE WORK

In this work, we introduce a novel framework to generate bilingual lexicon without any supervision. Different from existing models, we jointly transform the source and the target embedding space into a shared embedding space to capture the cross-lingual features of word embeddings. Based on the experimental results on eight language pairs, we show that the proposed model induces higher quality dictionaries than existing adversarial methods and even obtains best-published results on several language pairs. Besides, our method is robust to scenarios which are more realistic for low-resource languages.

In the future, we would like to extend our framework to multilingual scenarios, and apply our framework to other downstream applications such as unsupervised machine translation and domain adaptation.

ACKNOWLEDGMENT

he authors would like to thank the anonymous reviewers for their insightful comments. During this work, X. Bai was a M.S. candidate in Harbin Institute of Technology.

REFERENCES

[1] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, “Unsupervised neural machine translation,” in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Sy2ogebAW

[2] Z. Yang, W. Chen, F. Wang, and B. Xu, “Unsupervised neural machine translation with weight sharing,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 46–55. [Online]. Available: http://aclweb.org/anthology/P18-1005

[3] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, “Unsupervised machine translation using monolingual corpora only,” in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rkYTTf-AZ

[4] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, “Phrase-based & neural unsupervised machine translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 5039–5049. [Online]. Available: http://aclweb.org/anthology/D18-1549

[5] M. Artetxe, G. Labaka, and E. Agirre, “Unsupervised statistical machine translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Nov. 2018, pp. pp. 3632–3642. [Online]. Available: https://www.aclweb.org/anthology/D18-1399

[6] I. Vulić and M.-F. Moens, “Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings,” in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, New York, NY, USA, 2015, pp. 363–372. [Online]. Available: http://doi.acm.org/10.1145/2766462.2767752

[7] A. Klementiev, I. Titov, and B. Bhattarai, “Inducing crosslingual distributed representations of words,” in *Proc. COLING*, 2012, pp. 1459–1474. [Online]. Available: http://aclweb.org/anthology/C12-1089

[8] A. Mogadala and A. Rettinger, “Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 692–702. [Online]. Available: http://aclweb.org/anthology/N16-1083

[9] J. Guo, W. Che, D. Yarowsky, H. Wang, and T. Liu, “Cross-lingual dependency parsing based on distributed representations,” in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, Jul. 2015, pp. 1234–1244. [Online]. Available: http://www.aclweb.org/anthology/P15-1119

[10] S. Mayhew, C.-T. Tsai, and D. Roth, “Cheap translation for cross-lingual named entity recognition,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2536–2545. [Online]. Available: http://aclweb.org/anthology/D17-1269

[11] J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell, “Neural cross-lingual named entity recognition with minimal resources,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Nov. 2018, pp. 369–379. [Online]. Available: https://www.aclweb.org/anthology/D18-1034

[12] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” *CoRR*, vol. abs/1309.4168, 2013. [Online]. Available: http://arxiv.org/abs/1309.4168

[13] H. Aldarmaki, M. Mohan, and M. Diab, “Unsupervised word mapping using structural similarities in monolingual embeddings,” *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 185–196, Mar. 2018. [Online]. Available: http://aclweb.org/anthology/Q18-1014

[14] Y. Hoshen and L. Wolf, “Non-adversarial unsupervised word translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 469–478. [Online]. Available: http://aclweb.org/anthology/D18-1043

[15] M. Artetxe, G. Labaka, and E. Agirre, “A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 789–798. [Online]. Available: http://aclweb.org/anthology/P18-1073

[16] M. Zhang, Y. Liu, H. Luan, and M. Sun, “Adversarial training for unsupervised bilingual lexicon induction,” in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1959–1970. [Online]. Available: http://aclweb.org/anthology/P17-1179

[17] M. Zhang, Y. Liu, H. Luan, and M. Sun, “Earth mover’s distance minimization for unsupervised bilingual lexicon induction,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1934–1945. [Online]. Available: http://aclweb.org/anthology/D17-1207

[18] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H196sainb

[19] A. Søgaard, S. Ruder, and I. Vulić, “On the limitations of unsupervised bilingual dictionary induction,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 778–788. [Online]. Available: http://aclweb.org/anthology/P18-1072

[20] M. Faruqui and C. Dyer, “Improving vector space word representations using multilingual correlation,” in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 462–471. [Online]. Available: http://aclweb.org/anthology/E14-1049

[21] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2015, pp. 1006–1011. [Online]. Available: http://aclweb.org/anthology/N15-1104

[22] A. Lazaridou, G. Dinu, and M. Baroni, “Hubness and pollution: Delving into cross-space mapping for zero-shot learning,” in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Jul. 2015, pp. 270–280. [Online]. Available: http://www.aclweb.org/anthology/P15-1027

[23] Y. Zhang, D. Gaddy, R. Barzilay, and T. Jaakkola, “Ten pairs to tag—multilingual pos tagging via coarse mapping between embeddings,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1307–1317. [Online]. Available: http://aclweb.org/anthology/N16-1156

[24] M. Artetxe, G. Labaka, and E. Agirre, “Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations,” in *Proc. 32nd AAAI Conf.*, Feb. 2018, pp. 5012–5019. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16935

[25] N. Nakashole, “NORMA: Neighborhood sensitive maps for multilingual word embeddings,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Oct./Nov. 2018, pp. 512–522. [Online]. Available: https://www.aclweb.org/anthology/D18-1047

[26] Y. Doval, J. Camacho-Collados, L. Espinosa-Anke, and S. Schockaert, “Improving cross-lingual word embeddings by meeting in the middle,” in *Proc. Empirical Methods Natural Lang. Process.*, 2018, pp. 294–304. [Online]. Available: https://www.aclweb.org/anthology/D18-1027

[27] Y. Peirsman and S. Padó, “Cross-lingual induction of selectional preferences with bilingual vector spaces,” in *Proc. Human Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 921–929. [Online]. Available: http://aclweb.org/anthology/N10-1135

[28] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax,” in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=r1Aab85gg

[29] M. Artetxe, G. Labaka, and E. Agirre, “Learning bilingual word embeddings with (almost) no bilingual data,” in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 451–462. [Online]. Available: http://aclweb.org/anthology/P17-1042

[30] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[31] A. V. Miceli Barone, “Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders,” in *Proc. 1st Workshop Representation Learn. NLP*, 2016, pp. 121–126. [Online]. Available: http://aclweb.org/anthology/W16-1614

[32] X. Chen and C. Cardie, “Unsupervised multilingual word embeddings,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 261–270. [Online]. Available: http://aclweb.org/anthology/D18-1024

[33] Z.-Y. Dou, Z.-H. Zhou, and S. Huang, “Unsupervised bilingual lexicon induction via latent variable models,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 621–626. [Online]. Available: http://aclweb.org/anthology/D18-1062

[34] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proc. Int. Conf. Learn. Representations*, 2014.

[35] T. Mukherjee, M. Yamada, and T. Hospedales, “Learning unsupervised word translations without adversaries,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 627–632. [Online]. Available: http://aclweb.org/anthology/D18-1063

[36] R. Xu, Y. Yang, N. Otani, and Y. Wu, “Unsupervised cross-lingual transfer of word embedding spaces,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2465–2474. [Online]. Available: http://aclweb.org/anthology/D18-1268

[37] D. Alvarez-Melis and T. Jaakkola, "Gromov–Wasserstein alignment of word embedding spaces," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1881–1890. [Online]. Available: http://aclweb.org/anthology/D18-1214

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[39] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[40] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.,* 2014, pp. 1532–1543. [Online]. Available: http://aclweb.org/anthology/D14-1162

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[42] Y. Kementchedjhieva, S. Ruder, R. Cotterell, and A. Søgaard, "Generalizing procrustes analysis for better bilingual dictionary induction," in *Proc. 22nd Conf. Comput. Natural Lang. Learn.*, 2018, pp. 211–220. [Online]. Available: http://aclweb.org/anthology/K18-1021

[43] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu, "Deep multilingual correlation for improved word embeddings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2015, pp. 250–256. [Online]. Available: http://aclweb.org/anthology/N15-1028

[44] N. Nakashole and R. Flauger, "Characterizing departures from linearity in word translation," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 221–227. [Online]. Available: http://aclweb.org/anthology/P18-2036

[45] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2979–2984. [Online]. Available: http://aclweb.org/anthology/D18-1330

Xuefeng Bai received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2017, where he is currently working toward the M.S. degree . His research interests include machine translation and natural language processing.

Hailong Cao received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2006, where he is an Associate Professor with the School Computer Science and Technology. His research focusing on the teaching and research about natural language processing.

Kehai Chen received the B.S. degree from the Xi'an University of Technology, Xi'an, China, in 2010, the M.S. degree from University of Chinese Academy of Sciences, Beijing, China, in 2013, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2018. He was a internship researcher fellow with the National Institute of Information and Communications Technology, Japan in 2017. He has been a researcher with the National Institute of Information and Communications Technology, Japan since 2018. His research interests include machine translation and natural language processing.

Tiejun Zhao is a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include natural language understanding, content-based web information processing, and applied artificial intelligence. He has authored 3 academic books and 60 papers on journals and conferences in recent 3 years. He has been a PC member on ACL, COLING in current five years and was also assigned an MT Track Co-chair on COLING 2014.