# Unsupervised Bilingual Lexicon Induction via Latent Variable Models

Anonymous EMNLP submission

## Abstract

Bilingual lexicon extraction has been studied for decades and most previous methods have relied on parallel corpora or bilingual dictionaries. Recent studies have shown that it is possible to build a bilingual dictionary by aligning monolingual word embedding spaces in an unsupervised way. With the recent advances in generative models, we propose a novel approach which builds cross-lingual dictionaries via latent variable models and adversarial training with no parallel corpora. To demonstrate the effectiveness of our approach, we evaluate our approach on several language pairs and the experimental results show that our model could achieve competitive and even superior performance compared with several state-of-the-art models.

## 1 Introduction

Learning the representations of languages is a fundamental problem in natural language processing and most existing methods exploit the hypothesis that words occurring in similar contexts tend to have similar meanings (Pennington et al., 2014; Bojanowski et al., 2017), which could lead word vectors to capture semantic information. Mikolov et al. (2013) first point out that word embeddings learned on separate monolingual corpora exhibit similar structures. Based on this finding, they suggest it is possible to learn a linear mapping from a source to a target embedding space and then generate bilingual dictionaries. This simple yet effective approach has led researchers to investigate on improving cross-lingual word embeddings with the help of bilingual word lexicons (Faruqui and Dyer, 2014; Xing et al., 2015).

For low-resource languages and domains, cross-lingual signal would be hard and expensive to obtain, and thus it is necessary to reduce the need for bilingual supervision. Artetxe et al. (2017) successfully learn bilingual word embeddings with only a parallel vocabulary of aligned digits. Zhang et al. (2017) utilize adversarial training to obtain cross-lingual word embeddings without any parallel data. However, their performance is still significantly worse than supervised methods. By combining the merits of several previous works, Conneau et al. (2018) introduce a model that reaches and even outperforms supervised state-of-the-art methods with no parallel data.

In recent years, generative models have become more and more powerful. Both Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Autoencoders (VAEs) (Kingma and Welling, 2014) are prominent ones. In this work, we borrow the ideas from both GANs and VAEs to tackle the problem of bilingual lexicon induction. The basic idea is to learn latent variables that could capture semantic meaning of words, which would be helpful for bilingual lexicon induction. We also utilize adversarial training for our model and require no form of supervision. We evaluate our approach on several language pairs and experimental results demonstrate that our model could achieve promising performance. We further combine our model with several helpful techniques and show our model could perform competitively and even superiorly compared with several state-of-the-art methods.

## 2 Related Work

### 2.1 Bilingual Lexicon Induction

Extracting bilingual lexica has been studied by researchers for a long time. Mikolov et al. (2013) first observe there is isomorphic structure among word embeddings trained separately on monolingual corpora and they learn the linear transformation between languages. Zhang et al. (2016b) improve the method by constraining the transforma-

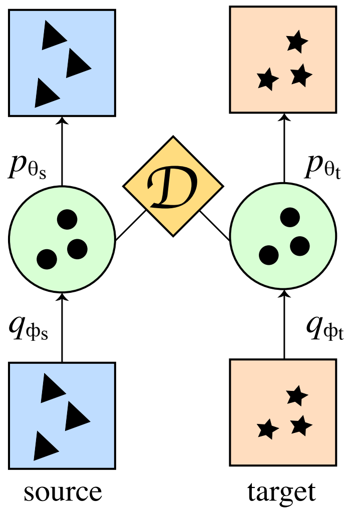

Figure 1: Illustration of our model. $\phi_s$ and $\phi_t$ map the source and target word embeddings into latent variables. Discriminator $D$ guides the two latent distributions to be the same.

tion matrix to be orthogonal. Xing et al. (2015) incorporate length normalization during training and maximize the cosine similarity instead. They point out that adding an orthogonality constraint can improve performance and has a closed-form solution, which was referred to as Procrustes approach in Smith et al. (2017). Canonical correlation analysis has also been used to map both languages to a shared vector space (Faruqui and Dyer, 2014; Lu et al., 2015).

To reduce the need for supervision signals, Artetxe et al. (2017) use identical digits and numbers to form an initial seed dictionary and then iteratively refine their results until convergence. Zhang et al. (2017) apply adversarial training to align monolingual word vector spaces with no supervision. Conneau et al. (2018) further improve the model by combining adversarial training and Procrustes approach, and their unsupervised approach could reach and even outperform state-of-the-art supervised approaches.

### 2.2 Generative Models

VAEs (Kingma and Welling, 2014) represent one of the most successful deep generative models. Standard VAEs assume observed variables are generated from latent variables and the latent variables are sampled from a simple Gaussian distribution. VAEs have been successfully applied in several natural language processing tasks before (Zhang et al., 2016a; Bowman et al., 2016).

GANs (Goodfellow et al., 2014) are another framework for estimating generative models via an adversarial process and have attracted huge attention. The basic strategy is to train a generative model and a discriminative model simultaneously via an adversarial process. Adversarial Autoencoder (Makhzani et al., 2015) is a probabilistic autoencoder that uses the GANs to perform variational inference. By combining a VAE with a GAN, Larsen et al. (2016) use learned feature representations in the GAN discriminator as the basis for the VAE reconstruction objective. GANs have been applied in machine translation before (Yang et al., 2018; Lample et al., 2018).

## 3 Proposed Approach

In this section, we first briefly introduce VAEs, and then we illustrate the details and training techniques of our proposed model.

### 3.1 Variational Autoencoder

Variational Autoencoders (VAEs) are deep generative model which are capable of learning complex density models for data via latent variables. Given a nonlinear generative model $p_\theta(x|z)$ with input $x \in \mathbb{R}^D$ and associated latent variable $z \in \mathbb{R}^L$ drawn from a prior distribution $p_0(z)$, the goal of VAEs is to use a recognition model, $q_\phi(z|x)$ to approximate the posterior distribution of the latent variables by maximizing the following variational lower bound

$$\mathcal{L}_{\theta,\phi} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x)||p_0(z)), \tag{1}$$

where KL refers to Kullback-Leibler divergence.

### 3.2 Our Model

Basically, our model assumes that the source word embedding $\{x_n\}$ and the target word embedding $\{y_n\}$ could be drawn from a same latent variable space $\{z_n\}$, where $\{z_n\}$ is capable of capturing semantic meaning of words.

In contrast to the standard VAE prior that assumes each latent embedding $z_n$ to be drawn from the same latent Gaussian, our model just requires the distribution of latent variables for source and target word embeddings to be equal. To achieve such a goal, we utilize adversarial training to guide the two latent distributions to match with each other.

As in adversarial training, we have networks $\phi_s$ and $\phi_t$ for both source and target space, striving to map words into the same latent space, while the discriminator $D$ is a binary classifier which tries to distinguish between the two languages. We also have reconstruction networks $\theta_s$ and $\theta_t$ as in VAEs.

The objective function for the discriminator $D$ could be formulated as

$$\begin{aligned}\mathcal{L}_D = &\ \mathbb{E}_{z_y \sim q_{\phi_t}(z|y)}[\log D(z_y)] \\ &+ \mathbb{E}_{z_x \sim q_{\phi_s}(z|x)}[\log(1 - D(z_x))].\end{aligned} \tag{2}$$

For the source side, the objective is to minimize

$$\begin{aligned}\mathcal{L}_{\phi_s,\theta_s} = &\ \mathbb{E}_{z_x \sim q_{\phi_s}(z|x)}[\log p_{\theta_s}(x|z_x)] \\ &- \mathbb{E}_{z_x \sim q_{\phi_s}(z|x)}[\log D(z_x)].\end{aligned} \tag{3}$$

Here we define $q_{\phi_s}(z|x) = \mathcal{N}(\mu_s(x), \Sigma_s(x))$, where $\mu_s(x) = W_{\mu_s}x$ and $\Sigma_s(x) = \exp(W_{\sigma_s}x)$; $W_{\mu_s}$ and $W_{\sigma_s}$ are learned parameters. We also define $p_{\theta_s}(x|z) = W_{\mu_s}^{\mathrm{T}}z$. The objective function and structure for $\phi_t$ are similar.

The basic framework of our model is shown in Figure 1. As we could see from the figure, our model tries to map the source and target word embedding into the same latent space which could capture the semantic meaning of words.

Theoretical analysis has revealed that adversarial training tries to minimize the Jensen-Shannon (JS) divergence between the real and fake distribution. Therefore, one can view our model as replace KL divergence in Equation 1 with JS divergence and change the Gaussian prior to the target distribution.

### 3.3 Training Strategy

Our model has two generators $\phi_s$ and $\phi_t$, and we have found that training them jointly would be extremely unstable. In this paper, we propose an iterative method to train our models. Basically, we first initialize $W_{\mu_t}$ to be identity matrix and train $\phi_s$ and $\theta_s$ on the source side. After convergence, we freeze $W_{\mu_s}$, and then train $\phi_t$ and $\theta_t$ in the target side. The pseudo-code for this process is shown in Algorithm 1. It should be noted that there is no variance once completing training.

## 4 Experiment

Our experiments could be divided into two parts. In the first part, we conduct experiments on small-scale datasets and our main baseline is Zhang et al. (2017). In the second part, we combine our model with several advanced techniques and we compare our model with Conneau et al. (2018) on large-scale datasets.

**Algorithm 1** Training Strategy

```
$W_{\mu_t} = I$
for $i = 1, \cdots, n_{iter}$ do
    while $\phi_s$ and $\theta_s$ have not converged do
        Update discriminator $D$
        Update $\phi_s$ and $\theta_s$
    end while
    while $\phi_t$ and $\theta_t$ have not converged do
        Update discriminator $D$
        Update $\phi_t$ and $\theta_t$
    end while
end for
```

| | | #tokens | vocab. size |
|---|---|---|---|
| zh-en | zh | 21m | 3,349 |
| | en | 53m | 5,154 |
| es-en | es | 61m | 4,774 |
| | en | 95m | 6,637 |
| it-en | it | 73m | 8,490 |
| | en | 93m | 6,597 |

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

### 4.1 Small-scale Datasets

In this section, our experiments focus on small-scale datasets and our main baseline model is adversarial autoencoder (Zhang et al., 2017). For justice, we use the same model selection strategy with Zhang et al. (2017), i.e. we choose the model whose sum of reconstruction loss and classification accuracy is the least. Performance is measured by top-1 accuracy.

#### 4.1.1 Experiments on Chinese-English Dataset

For this set of experiments, we use the same data as Zhang et al. (2017). The statistics of the final training data is given in Table 1. We use Chinese-English Translation Lexicon Version 3.0 (LDC2002L27) as our ground truth bilingual lexicon for evaluation.

The baseline models are MonoGiza system (Dou et al., 2015), translation matrix (TM) (Mikolov et al., 2013), isometric alignment (IA) (Zhang et al., 2016b) and adversarial training approach (Zhang et al., 2017).

Table 2 summarizes the performance of baseline models and our approach. The results of baseline models are cited from Zhang et al. (2017). As we can see from the table, our model could achieve superior performance compared with other baseline models. Table 3 lists some word translation examples given by our model.

| Model | #seeds | Accuracy (%) |
|---|---|---|
| MonoGiza w/o emb. | 0 | 0.05 |
| MonoGiza w/ emb. | 0 | 0.09 |
| TM | 50 | 0.29 |
| IA | 100 | 21.79 |
| Zhang et al. (2017) | 0 | 43.31 |
| Ours | 0 | **51.37** |

Table 2: Experimental results on Chinese-English dataset.

| 航空 | 铁路 | 时代 | 学校 |
|---|---|---|---|
| airline | rail | antiquity | **school** |
| **aviation** | **railway** | **era** | education |
| airliner | **railroad** | century | college |
| service | freight | medieval | student |
| flight | metro | historian | teacher |

Table 3: Word translation examples for Chinese-English dataset. Ground truths are marked in bold.

### 4.1.2 Experiments on Other Language Pairs Datasets

We also conduct experiments on Spanish-English and Italian-English language pairs. Again, we use the same dataset with Zhang et al. (2017). and the statistics are shown in Table 1. The ground truth bilingual lexica for Spanish-English and Italian-English are obtained from Multilingual Unsupervised and Supervised Embeddings (MUSE) [1].

| | Model | Accuracy (%) |
|---|---|---|
| es-en | Zhang et al. (2017) | 69.22 |
| | Ours | **75.21** |
| it-en | Zhang et al. (2017) | 55.31 |
| | Ours | **61.08** |

Table 4: Experimental results on Spanish-English and Italian-English datasets.

The experimental results are shown in Table 4. Because Spanish, Italian and English are closely related languages, the accuracy would be higher than the Chinese-English dataset. Our model is able to outperform baseline model in this setting.

| Model | Accuracy (%) | | |
|---|---|---|---|
| | en-es | en-ru | en-zh |
| *Methods without refinement* | | | |
| Adv-NN | 69.8 | 29.1 | 18.5 |
| Adv-CSLS | 75.7 | 37.2 | 23.4 |
| Ours-NN | 71.8 | 32.8 | 22.9 |
| Ours-CSLS | **76.6** | **39.3** | **26.0** |
| *Methods with refinement* | | | |
| Adv-Refine-NN | 79.1 | 37.3 | 30.9 |
| Adv-Refine-CSLS | 81.7 | 44.0 | 32.5 |
| Ours-Refine-NN | 79.1 | 42.7 | 32.5 |
| Ours-Refine-CSLS | **82.1** | **48.7** | **33.3** |

Table 5: Experimental results on large-scale datasets. Language codes: en=English, es = Spanish, ru = Russian, zh = Chinese.

## 4.2 Large-scale Datasets

In this section, we integrate our method with Conneau et al. (2018). We replace their first step, namely the adversarial training step, with our model. Basically, we first map the source and target embeddings into the latent space using our algorithm, and then fine-tune the identity mapping in the latent space with the closed-form Procrustes solution. We use their similarity measure, namely cross-domain similarity local scaling (CSLS), to produce reliable matching pairs and validation criterion for unsupervised model selection.

We conduct experiments on English-Spanish, English-Russian and English-Chinese datasets, which are the same as Conneau et al. (2018). The results are shown in Table 5. As seen, our model could consistently achieve better performance compared with adversarial training. After refinement, our model could further achieve competitive and even superior results compared with state-of-the-art unsupervised methods. This further demonstrates the capacity of our model.

## 5 Conclusion

Based on the assumption that word vectors in different languages could be drawn from a same latent variable space, we propose a novel approach which builds cross-lingual dictionaries via latent variable models and adversarial training with no parallel corpora. Experimental results on several language pairs have demonstrated the effectiveness and universality of our model. We believe our method could advance the field and help other areas such as unsupervised machine translation.

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

## References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. In *International Joint Conference on Natural Language Processing (IJCNLP)*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Conference on Neural Information Processing Systems (NIPS)*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. *International Conference on Learning Representations (ICLR)*.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning (ICML)*.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations (ICLR)*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Improving neural machine translation with conditional sequence generative adversarial nets. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Biao Zhang, Deyi Xiong, Hong Duan, Min Zhang, et al. 2016a. Variational neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016b. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.