A Relaxed Matching Procedure for Unsupervised BLI

Xu Zhao[1], **Zihao Wang**[1], **Hao Wu**[2], **Zhang Yong**[1] †
[1]BNRist, Department of Computer Science and Technology, RIIT,
Institute of Internet Industry, Tsinghua University, Beijing, China
[2]Department of Mathematical Sciences, Tsinghua University, Beijing, China
{zhaoxu18, wzh17}@mails.tsinghua.edu.cn
{hwu, zhangyong05}@tsinghua.edu.cn

Abstract

Recently unsupervised **Bilingual Lexicon Induction(BLI)** without any parallel corpus has attracted much research interest. One of the crucial parts in methods for the BLI task is the matching procedure. Previous works impose a too strong constraint on the matching and lead to many counterintuitive translation pairings. Thus, We propose a relaxed matching procedure to find a more precise matching between two languages. We also find that aligning source and target language embedding space bidirectionally will bring significant improvement. We follow the previous iterative framework to conduct experiments. Results on standard benchmark demonstrate the effectiveness of our proposed method, which substantially outperforms previous unsupervised methods.

1 Introduction

Pretrained word embeddings (Mikolov et al., 2013b) are the basis of many other natural language processing and machine learning systems. Word embeddings of a specific language contain rich syntax and semantic information. Mikolov et al. (2013a) stated that the continuous embedding spaces exhibit similar structures across different languages, and we can exploit the similarity by a linear transformation from source embedding space to target embedding space. This similarity derives the **Bilingual Lexicon Induction(BLI)** task. The goal of bilingual lexicon induction is to align two languages' embedding space and generates word translation lexicon automatically. This fundamental problem in natural language processing benefits much other research such as sentence translation (Rapp, 1995; Fung, 1995), unsupervised machine translation (Lample et al., 2017), cross-lingual information retrieval (Lavrenko et al., 2002).

Recent endeavors (Lample et al., 2018; Alvarez-Melis and Jaakkola, 2018; Grave et al., 2019; Artetxe et al., 2017) have proven that unsupervised BLI's performance is even on par with the supervised methods. A crucial part of these approaches is the **matching procedure**, i.e., how to generate the translation plan. Alvarez-Melis and Jaakkola (2018) used Gromov-Wasserstein distance to approximate the matching between languages. Grave et al. (2019) regarded it as a classic optimal transport problem and used the sinkhorn algorithm (Cuturi, 2013) to compute the translation plan.

In this work, we follow the previous iterative framework but use a different matching procedure. Previous iterative algorithms required to compute an approximate 1 to 1 matching every step. This 1 to 1 constraint brings out many redundant matchings. Thus in order to avoid this problem, we relax the constraint and control the relaxation degree by adding two KL divergence regularization terms to the original loss function. This relaxation derives a more precise matching and significantly improves performance. Then we propose a bidirectional optimization framework to optimize the mapping from source to target and from target to source simultaneously. In the section of experiments, we verify the effectiveness of our method, and results show our method outperforms many SOTA methods on the BLI task.

2 Background

The early works for the BLI task require a parallel lexicon between languages. Given two embedding matrices X and Y with shape $n \times d$ (n:word number, d:vector dimension) of two languages and word x_i in X is the translation of word y_i in Y, i.e., we get a parallel lexicon $X \rightarrow Y$. Mikolov et al. (2013a) pointed out that we could exploit the similarities of monolingual embedding spaces by learning a linear transformation $W^\star$ such that

$$W^\star = \underset{W \in M_d(\mathbb{R})}{\arg\min} \|XW - Y\|_F^2 \qquad (1)$$

† Yong Zhang is the corresponding author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3036–3041
July 5 - 10, 2020.

where $M_d(\mathbb{R})$ is the space of $d \times d$ matrices of real numbers. Xing et al. (2015) stated that enforcing an orthogonal constraint on W would improve performance. There is a closed-form solution to this problem called **Procrutes**: $W^\star = Q = UV^T$ where $USV^T = XY^T$.

Under the unsupervised condition without parallel lexicon, i.e., vectors in X and Y are totally out of order, Lample et al. (2018) proposed a domain-adversarial approach for learning $W^\star$. On account of the ground truth that monolingual embedding spaces of different languages keep similar spatial structures, Alvarez-Melis and Jaakkola (2018) applied the Gromov-Wasserstein distance based on infrastructure to find the corresponding translation pairings between X and Y and further derived the orthogonal mapping Q. Grave et al. (2019) formulated the unsupervised BLI task as

$$\min_{Q \in \mathcal{O}_d, P \in \mathcal{P}_n} \|XQ - PY\|_F^2 \tag{2}$$

where $\mathcal{O}_d$ is the set of orthogonal matrices and $\mathcal{P}_n$ is is the set of permutation matrices.Given Q, estimating P in Problem (2) is equivalent to the minimization of the 2-Wasserstein distance between the two sets of points: XQ and Y.

$$W_2^2(XQ, Y) = \min_{P \in \mathcal{P}_n} \langle D, P \rangle \tag{3}$$

where $D_{ij} = \|x_i Q - y_j\|_2^2$ and $\langle D, P \rangle = \sum_{i,j} P_{ij} D_{ij}$ denotes the matrix inner product. Grave et al. (2019) proposed a stochastic algorithm to estimate Q and P jointly. Problem (3) is the standard optimal transport problem that can be solved by Earth Mover Distance linear program with $O(n^3)$ time complexity. Considering the computational cost, Zhang et al. (2017) and Grave et al. (2019) used the Sinkhorn algorithm (Cuturi, 2013) to estimate P by solving the entropy regularized optimal tranpsort problem (Peyré et al., 2019).

We also take Problem (2) as our loss function and our model shares a similar alternative framework with Grave et al. (2019). However, we argue that the permutation matrix constraint on P is too strong, which leads to many inaccurate and redundant matchings between X and Y, so we relax it by unbalanced optimal transport.

Alaux et al. (2019) extended the line of BLI to the problem of aligning multiple languages to a common space. Zhou et al. (2019) estimated Q by a density matching method called normalizing flow. Artetxe et al. (2018) proposed a multi-step framework of linear transformations that generalizes a substantial body of previous work. Garneau et al. (2019) further investigated the robustness of Artetxe et al. (2018)'s model by introducing four new languages that are less similar to English than the ones proposed by the original paper. Artetxe et al. (2019) proposed an alternative approach to this problem that builds on the recent work on unsupervised machine translation.

3 Proposed Method

In this section, we propose a method for the BLI task. As mentioned in the background, we take Problem (2) as our loss function and use a similar optimization framework in Grave et al. (2019) to estimate P and Q alternatively. Our method focuses on the estimation of P and tries to find a more precise matching P between XQ and Y. Estimation of Q is by stochastic gradient descent. We also propose a bidirectional optimization framework in section 3.2.

3.1 Relaxed Matching Procedure

Regarding embedding set X and Y as two discrete distributions $\mu = \sum_{i=1}^{I} u_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{J} v_j \delta_{y_j}$, where u (or v) is column vector satisfies $\sum_i u_i = 1, u_i > 0$ (v is similar), δ_x is the Dirac function supported on point x.

Standard optimal transport enforces the optimal transport plan to be the joint distribution $P \in \mathcal{P}_n$. This setting leads to the result that every mass in μ should be matched to the same mass in ν. Recent application of unbalanced optimal transport (Wang et al., 2019) shows that the relaxation of the marginal condition could lead to more flexible and local matching, which avoids some counterintuitive matchings of source-target mass pairs with high transportation cost.

The formulation of unbalanced optimal transport (Chizat et al., 2018a) differs from the balanced optimal transport in two ways. Firstly, the set of transport plans to be optimized is generalized to $\mathbb{R}_+^{I \times J}$. Secondly, the marginal conditions of the Problem (3) are relaxed by two KL-divergence terms.

$$\begin{aligned} \min_{P \in \mathbb{R}_+^{I \times J}} \langle D, P \rangle &+ \lambda_1 \mathcal{KL}(P \mathbb{1}_J || u) \\ &+ \lambda_2 \mathcal{KL}(P^T \mathbb{1}_I || v) \end{aligned} \tag{4}$$

where $\mathcal{KL}(p||q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) - p_i + q_i$ is the KL divergence.

Algorithm 1 Generalized Sinkhorn Algorithm

Require: source and target measure $\mu_i \in \mathbb{R}_+^m, \nu_j \in \mathbb{R}^n$, entropy regularizer ϵ, KL relaxation coefficient λ_1, λ_2 and distance matrixD_{ij}.

Ensure: Transport PlanP_{ij}

Initialize $u \leftarrow 0 \in \mathbb{R}^m$, $v \leftarrow 0 \in \mathbb{R}^n$, $K \leftarrow e^{-D/\gamma} \in \mathbb{R}^{m \times n}$
while not converge **do**
  $u \leftarrow \left(\frac{\mu}{Kv}\right)^{\frac{\lambda_1}{\epsilon+\lambda_1}}$
  $v \leftarrow \left(\frac{\nu}{K^\top u}\right)^{\frac{\lambda_2}{\epsilon+\lambda_2}}$
end while
$P \leftarrow \text{diag}(u) K \text{diag}(v)$

We estimate P by considering the relaxed Problem (4) instead of the original Problem (3) in (Grave et al., 2019). Problem (4) could also be solved by entropy regularization with the generalized Sinkhorn algorithm (Chizat et al., 2018b; Wang et al., 2019; Peyré et al., 2019).

In short, we already have an algorithm to obtain the minimum of the Problem (4). In order to avoid the hubness phenomenon, we replace l_2 distance of embedding with the $rcsls$ distance proposed in Joulin et al. (2018) formalized as $D_{ij} = rcsls(x_i Q, y_j)$. $rcsls$ can not provide significantly better results than euclidean distance in our evaluation. However, previous study suggests that RCSLS could be considered as a better metric between words than euclidean distance. So we propose our approach with RCSLS. The "relaxed matching" procedure and the "bi-directional optimization" we proposed bring most of the improvement.

We call this relaxed estimation of P as **Relaxed Matching Procedure(RMP)**. With RMP only when two points are less than some radius apart from each other, they may be matched together. Thus we can avoid some counterintuitive matchings and obtain a more precise matching P. In the section of experiments we will verify the effectiveness of RMP.

3.2 Bidirectional Optimization

Previous research solved the mapping X to Y and the mapping Y to X as two independent problems, i.e., they tried to learn two orthogonal matrix Q_1 and Q_2 to match the XQ_1 with Y and YQ_2 with X, respectively. Intuitively from the aspect of point cloud matching, we consider these two problems in opposite directions are symmetric. Thus we propose an optimization framework to solve only one Q for both directions.

Algorithm 2 Bidirectional Optimization with RMP

Require: word vectors from two languagesX, Y

Ensure: Transformation Q

for each $e \in [1, E]$ **do**
  for each $i \in [1, I]$ **do**
    Draw X_b, Y_b of size b from X and Y
    set $rand = random()$
    if $rand \mod 2 = 1$ **then**
      $Y_b, X_b, Q \Leftarrow X_b, Y_b, Q^T$
    end if
    Run RMP by solving Problem (4) and obtain P^*
    Update Q by gradient descent and Procrutes
    if $rand \mod 2 = 1$ **then**
      $Q \Leftarrow Q^T$
    end if
  end for
end for

In our approach, we match XQ with Y and YQ^T with X simultaneously. Based on the stochastic optimization framework of Grave et al. (2019), we randomly choose one direction to optimize at each iteration.

The entire process of our method is summarized in Algorithm 2. At iteration i, we start with sampling batches X_b, Y_b with shape $\mathbb{R}^{b \times d}$. Then we generate a random integer $rand$ and choose to map X_bQ to Y_b or map Y_bQ^T to X_b by $rand$'s parity. Given the mapping direction, we run the RMP procedure to solve Problem (4) by sinkhorn and obtain a matching matrix P^* between X_bQ and Y_b(or Y_bQ^T and X). Finally we use gradient descent and procrutes to update Q by the given P^*. The procedure of Q's update is detailed in Grave et al. (2019).

4 Experiments

In this section, we evaluate our method in two settings. First, We conduct distillation experiments to verify the effectiveness of RMP and bidirectinal optimization. Then we compare our method consisting of both RMP and bi-directional optimization with various SOTA methods on the BLI task.

DataSets[*] We conduct word translation experiments on 6 pairs of languages and use pretrained

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

Method		EN-ES		EN-FR		EN-DE		EN-RU		EN-IT		Avg.
	Supervision	→	←	→	←	→	←	→	←	→	←	
Proc.	5K words	81.9	83.4	82.1	82.4	74.2	72.7	51.7	63.7	77.4	77.9	74.7
RCSLS	5K words	84.1	86.3	83.3	84.1	79.1	76.3	57.9	67.2			77.3
GW	None	81.7	80.4	81.3	78.9	71.9	**78.2**	45.1	43.7	78.9	75.2	71.5
Adv. - Refine	None	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	77.9	77.5	73.4
W.Proc. - Refine	None	**82.8**	84.1	82.6	82.9	75.4	73.3	43.7	59.1			73.0
Dema - Refine	None	82.8	84.9	82.6	82.4	75.3	74.9	46.9	62.4			74.0
Ours - Refine	None	82.7	**85.8**	**83.0**	**83.8**	**76.2**	74.9	**48.1**	**64.7**	**79.1**	**80.3**	**75.9**

Table 1: Comparison between SOTA methods on BLI task. Methods in Line 1-2 are supervised. Methods in Line 3-8 are unsupervised. Except the GW method, other unsupervised methods are refined. In bold, the best among unsupervised approaches. All numbers of others are taken from their papers. ('EN': English, 'ES': Spanish, 'FR': French, 'DE': German, 'RU': Russian, 'IT': Italian).

word embedding from fasttext. We use the bilingual dictionaries opensourced in the work (Lample et al., 2018) as our evaluate set.We use the CSLS retrieval method for evaluation as Lample et al. (2018) in both settings. All the translation accuracy reported is the precision at 1 with CSLS criterion. We open the source code on Github[†].

4.1 Main Results

Through the experimental evaluation, we seek to demonstrate the effectiveness of our method compared to other SOTA methods. The word embeddings are normalized and centered before entering the model. We start with a batch size 500 and 2000 iterations each epoch. We double the batch size and quarter the iteration number after each epoch. First 2.5K words are taken for initialization, and samples are only drawn from the first 20K words in the frequently ranking vocabulary. The coefficients λ_1 and λ_2 of the relaxed terms in Problem (4) are both set to 0.001.

Baselines We take basic Procrutes and RCSLS-Loss of Joulin et al. (2018) as two supervised baselines. Five unsupervised methods are also taken into accounts: the Gromov Wasserstein matching method of Alvarez-Melis and Jaakkola (2018), the adversarial training(Adv.-Refine) of Lample et al. (2018), the Wasserstein Procrutes method(W.Proc.-Refine) of Grave et al. (2019), the density matching method(Dema-Refine) of Zhou et al. (2019).

In Table 1, it's shown that leading by an average of 2 percentage points, our approach outperforms other unsupervised methods in most instances and is on par with the supervised method on some language pairs. Surprisingly we find that our method achieves significant progress in some tough cases such as English - Russian, English - Italian, which contain lots of noise. Our method guarantees the precision of mapping computed every step which achieves the effect of noise reduction.

However, there still exists an noticeable gap between our method and the supervised RCSLS method, which indicates further research can be conducted to absorb the superiority of this metric to unsupervised methods.

We also compare our method with W.Proc on two non-English pairs including FR-DE and FR-ES to show how bidirectional relaxed matching improves the performance and results are presented in Table 2. Most of the recent researches didn't report results of non-English pairs, which makes it hard for fair comparison. However from the results in Table 2, we could find that our method keeps an advantage over W.Proc. Note that the W.Proc. results here are our implementation rather than that are reported in the original paper.

[†]https://github.com/BestActionNow/bidirectional-RMP

	FR-DE	DE-FR	FR-ES	ES-FR
W.Proc.	65.8	73.5	82.0	84.9
Ours-Refine	67.7	74.0	83.3	84.9

Table 2: Comparision bewtween W.Proc. and our method on non-English language pairs

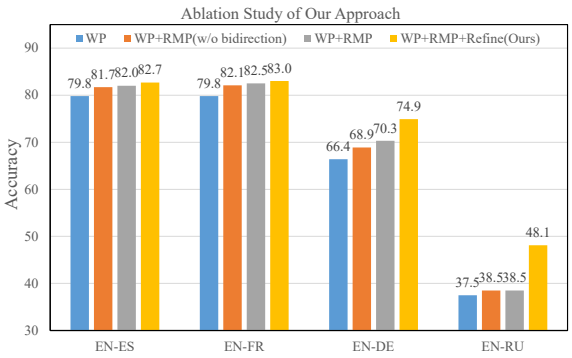

Figure 1: Ablation study of our methods' effectiveness. 'WP' refers to the original Wasserstein Procrutes Method proposed by Grave et al. (2019). 'WP-RMP' applies RMP to 'WP'. 'WP-RMP-bidiretion' applies bidirectional optimization framework to 'WP-RMP'. 'WP-RMP-bidirection-refine' applies the refinement procedure to 'WP-RMP-bidirection'.('EN': English, 'ES': Spanish, 'FR': French, 'DE': German, 'RU': Russian, 'IT': Italian).

4.2 Ablation Study

The algorithms for BLI could be roughly divided into three parts: 1. initialization, 2 iterative optimization, and 3. refinement procedure, such as Lample et al. (2017). W.Proc.(Grave et al., 2019) only covers the first two parts. Our approaches, i.e. relaxed matching and bi-directional optimization are categorized into the second part. To ensure a fair comparison, W.Proc.-Refine is compared to ours-Refine which is discussed in next section. To verify the effectiveness of RMP and bidirectional optimization directly, we apply them to the method proposed in Grave et al. (2019) one by one. We take the same implementation and hyperparameters reported in their paper and code [‡] but using RMP to solve P instead of ordinary 2-Wasserstein.

On four language pairs, We applied RMP, bidirectional optimization and refinement procedure to original W.Proc. gradually and evaluate the performance change. In Figure 1 it's clearly shown that after applying bidirectional RMP, the translation accuracy improves by 3 percentage averagely. The results of 'WP-RMP' are worse than 'WP-RMP-bidirection' but better than original 'WP'. Moreover, we find that by applying RMP, a more precise P not only eliminates many unnecessary matchings but also leads to a faster converge of the optimization procedure. Furthurmore, the effectiveness of refinement procedure is quite significant.

To summarize, we consider the average of scores (from en-es to ru-en). By mitigating the counter-intuitive pairs by polysemies and obscure words, the "relaxed matching" procedure improves the average score about 2 points, the "bi-directional optimization" improves the average score about 0.6 points. From the results we could get some inspiration that our ideas of relaxed matching and bidirectional optimization can also be applied to other frameworks such as adversarial training by Lample et al. (2017) and Gromov-Wasserstein by Alvarez-Melis and Jaakkola (2018).

[‡]https://github.com/facebookresearch/fastText/alignment

5 Conclusion

This paper focuses on the matching procedure of BLI task. Our key insight is that the relaxed matching mitigates the counter-intuitive pairs by polysemy and obscure words, which is supported by comparing W.Proc.-RMP with W.Proc in Table 1. The optimal transport constraint considered by W.Proc. is not proper for BLI tasks. Moreover, Our approach also optimizes the translation mapping Q in a bi-directional way, and has been shown better than all other unsupervised SOTA models with the refinement in Table 1.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (11871297, 91646202), National Key R&D Program of China(2018YFB1404401, 2018YFB1402701), Tsinghua University Initiative Scientific Research Program.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. CoRR, abs/1811.01124.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 1881–1890.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In Thirty-Second AAAI Conference on Artificial Intelligence.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 5002–5007.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. 2018a. An interpolating distance between optimal transport and fisher–rao metrics. Foundations of Computational Mathematics, 18(1):1–44.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. 2018b. Scaling algorithms for unbalanced optimal transport problems. Mathematics of Computation, 87(314):2563–2609.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pages 2292–2300.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In Third Workshop on Very Large Corpora.

Nicolas Garneau, Mathieu Godbout, David Beauchemin, Audrey Durand, and Luc Lamontagne. 2019. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings: Making the method robustly reproducible as well. CoRR, abs/1912.01706.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan, pages 1880–1890.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 2979–2984.

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

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.

Victor Lavrenko, Martin Choquette, and W Bruce Croft. 2002. Cross-lingual relevance models. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 175–182. ACM.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport. Foundations and Trends® in Machine Learning, 11(5-6):355–607.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. arXiv preprint cmp-lg/9505037.

Zihao Wang, Datong Zhou, Yong Zhang, Hao Wu, and Chenglong Bao. 2019. Wasserstein-fisher-rao document distance. CoRR, abs/1904.10294.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, pages 1006–1011.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth movers distance minimization for unsupervised bilingual lexicon induction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1934–1945.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 1588–1598.