

Duality Regularization for Unsupervised Bilingual Lexicon Induction

Xuefeng Bai^{1,2,3}, Yue Zhang^{2,3}, Hailong Cao⁴, Tiejun Zhao⁴

¹Zhejiang University

²School of Engineering, Westlake University

³Institute of Advanced Technology, Westlake Institute for Advanced Study

⁴Harbin Institute of Technology

xfbai.hk@gmail.com, zhangyue@westlake.edu.cn
caohailong2008@gmail.com, tzhao@hit.edu.cn

Abstract

Unsupervised bilingual lexicon induction naturally exhibits duality, which results from symmetry in back-translation. For example, EN-IT and IT-EN induction can be mutually primal and dual problems. Current state-of-the-art methods, however, consider the two tasks independently. In this paper, we propose to train primal and dual models jointly, using regularizers to encourage consistency in back translation cycles. Experiments across 6 language pairs show that the proposed method significantly outperforms competitive baselines, obtaining the best published results on a standard benchmark.

1 Introduction

Unsupervised bilingual lexicon induction (UBLI) has been shown to benefit NLP tasks for low resource languages, including unsupervised NMT (Artetxe et al., 2018b,c; Yang et al., 2018; Lample et al., 2018a,b), information retrieval (Vulić and Moens, 2015; Litschko et al., 2018), dependency parsing (Guo et al., 2015), and named entity recognition (Mayhew et al., 2017; Xie et al., 2018).

Recent research has attempted to induce unsupervised bilingual lexicons by aligning monolingual word vector spaces (Zhang et al., 2017a; Conneau et al., 2018; Aldarmaki et al., 2018; Artetxe et al., 2018a; Alvarez-Melis and Jaakkola, 2018; Mukherjee et al., 2018). Given a pair of languages, their word alignment is inherently a bi-directional problem (e.g. English-Italian vs Italian-English). However, most existing research considers mapping from one language to another without making use of symmetry. Our experiments show that separately learned UBLI models are not always consistent in opposite directions. As shown in Figure 1a, when the model of Conneau et al. (2018) is applied to English and

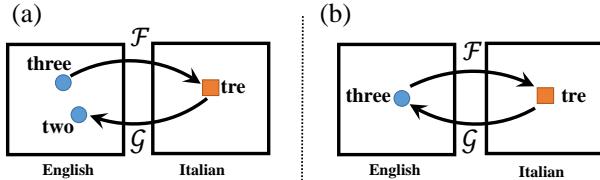


Figure 1: (a) Inconsistency between primal model \mathcal{F} and the dual model \mathcal{G} . (b) An ideal scenario.

Italian, the primal model maps the word “three” to the Italian word “tre”, but the dual model maps “tre” to “two” instead of “three”.

We propose to address this issue by exploiting duality, encouraging forward and backward mappings to form a closed loop (Figure 1b). In particular, we extend the model of Conneau et al. (2018) by using a cycle consistency loss (Zhou et al., 2016) to regularize two models in opposite directions. Experiments on two benchmark datasets show that the simple method of enforcing consistency gives better results in both directions. Our model significantly outperforms competitive baselines, obtaining the best published results. We release our code at xxx.

2 Related Work

UBLI. A typical line of work uses adversarial training (Miceli Barone, 2016; Zhang et al., 2017a,b; Conneau et al., 2018), matching the distributions of source and target word embeddings through generative adversarial networks (Goodfellow et al., 2014). Non-adversarial approaches have also been explored. For instance, Mukherjee et al. (2018) use squared-loss mutual information to search for optimal cross-lingual word pairing. Artetxe et al. (2018a) and Hoshen and Wolf (2018) exploit the structural similarity of word embedding spaces to learn word mappings. In this paper, we choose Conneau et al. (2018) as our baseline as it is theoretically attractive and gives strong results on large-scale datasets.

Cycle Consistency. Forward-backward consistency has been used to discover the correspondence between unpaired images (Zhu et al., 2017; Kim et al., 2017). In machine translation, similar ideas were exploited, He et al. (2016), Xia et al. (2017) and Wang et al. (2018) use dual learning to train two opposite language translators by minimizing the reconstruction loss. Sennrich et al. (2016) consider back-translation, where a backward model is used to build synthetic parallel corpus and a forward model learns to generate genuine text based on the synthetic output.

Closer to our method, Chandar et al. (2014) jointly train two autoencoders to learn supervised bilingual word embeddings. Xu et al. (2018) use sinkhorn distance (Cuturi, 2013) and back-translation to align word embeddings. However, they cannot perform fully unsupervised training, relying on WGAN (Arjovsky et al., 2017) for providing initial mappings. Concurrent with our work, Mohiuddin and Joty (2019) build a adversarial autoencoder with cycle consistency loss and post-cycle reconstruction loss. In contrast to these works, our method is fully unsupervised, simpler, and empirically more effective.

3 Approach

We take Conneau et al. (2018) as our baseline, introducing a novel regularizer to enforce cycle consistency. Let $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$ be two sets of n and m word embeddings for a source and a target language, respectively. The primal UBLI task aims to learn a linear mapping $\mathcal{F} : X \rightarrow Y$ such that for each x_i , $\mathcal{F}(x_i)$ corresponds to its translation in Y . Similarly, a linear mapping $\mathcal{G} : Y \rightarrow X$ is defined for the dual task. In addition, we introduce two language discriminators D_x and D_y , which are trained to discriminate between the mapped word embeddings and the original word embeddings.

3.1 Baseline Adversarial Model

Conneau et al. (2018) align two word embedding spaces through generative adversarial networks, in which two networks are trained simultaneously. Specifically, take the primal UBLI task as an example, the linear mapping \mathcal{F} tries to generate “fake” word embeddings $\mathcal{F}(x)$ that look similar to word embeddings from Y , while the discriminator D_y aims to distinguish between “fake” and real word embeddings from Y . Formally, this

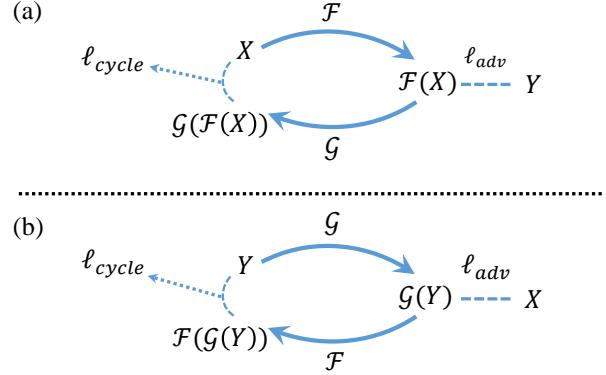


Figure 2: The proposed framework. (a) $X \rightarrow \mathcal{F}(X) \rightarrow \mathcal{G}(\mathcal{F}(X)) \rightarrow X$; (b) $Y \rightarrow \mathcal{G}(Y) \rightarrow \mathcal{F}(\mathcal{G}(Y)) \rightarrow Y$.

idea can be expressed as the minmax game $\min_{\mathcal{F}} \max_{D_y} \ell_{adv}(\mathcal{F}, D_y, X, Y)$, where

$$\begin{aligned} \ell_{adv}(\mathcal{F}, D_y, X, Y) = & \frac{1}{m} \sum_{j=1}^m \log P_{D_y}(\text{src} = 1 | y_j) \\ & + \frac{1}{n} \sum_{i=1}^n \log P_{D_y}(\text{src} = 0 | \mathcal{F}(x_i)). \end{aligned} \quad (1)$$

$P_{D_y}(\text{src} | y_j)$ is a model probability from D_y to distinguish whether word embedding y_j is coming from the target language ($\text{src} = 1$) or the primal mapping \mathcal{F} ($\text{src} = 0$). Similarly, the dual UBLI problem can be formulated as $\min_{\mathcal{G}} \max_{D_x} \ell_{adv}(\mathcal{G}, D_x, Y, X)$, where \mathcal{G} is the dual mapping, and D_x is a source discriminator.

Theoretically, a unique solution for above minmax game exists, with the mapping and the discriminator reaching a nash equilibrium. Since the adversarial training happens at the distribution level, no cross-lingual supervision is required.

3.2 Regularizers for Dual Models

We train \mathcal{F} and \mathcal{G} jointly and introduce two regularizers. Formally, we hope that $\mathcal{G}(\mathcal{F}(X))$ is similar to X and $\mathcal{F}(\mathcal{G}(Y))$ is similar to Y . We implement this constraint as a cycle consistency loss. As a result, the proposed model has two learning objectives: i) an adversarial loss (ℓ_{adv}) for each model as in the baseline. ii) a cycle consistency loss (ℓ_{cycle}) on each side to avoid \mathcal{F} and \mathcal{G} from contradicting each other. The overall architecture of our model is illustrated in Figure 2.

Cycle Consistency Loss. We introduce

$$\begin{aligned} \ell_{cycle}(\mathcal{F}, \mathcal{G}, X) = & \frac{1}{n} \sum_{i=1}^n \Delta(x_i, \mathcal{G}(\mathcal{F}(x_i))), \\ \ell_{cycle}(\mathcal{F}, \mathcal{G}, Y) = & \frac{1}{m} \sum_{j=1}^m \Delta(y_j, \mathcal{F}(\mathcal{G}(y_j))), \end{aligned} \quad (2)$$

where Δ denotes the discrepancy criterion, which is set as the average cosine similarity in our model.

Full objective. The final objective is:

$$\begin{aligned} \ell(\mathcal{F}, \mathcal{G}, D_x, D_y, X, Y) = & \\ & \ell_{adv}(\mathcal{F}, D_y, X, Y) + \ell_{adv}(\mathcal{G}, D_x, Y, X) \\ & + \ell_{cycle}(\mathcal{F}, \mathcal{G}, X) + \ell_{cycle}(\mathcal{F}, \mathcal{G}, Y). \quad (3) \end{aligned}$$

3.3 Model Selection

We follow Conneau et al. (2018), using an unsupervised criterion to perform model selection. In preliminary experiments, we find in adversarial training that the single-direction criterion $S(\mathcal{F}, X, Y)$ by Conneau et al. (2018) does not always work well. To address this, we make a simple extension by calculating the weighted average of forward and backward scores:

$$S_a = \lambda S(\mathcal{F}, X, Y) + (1 - \lambda) S(\mathcal{G}, X, Y), \quad (4)$$

Where λ is a hyperparameter to control the importance of the two objectives.¹ Here S first generates bilingual lexicons by learned mappings, and then computes the average cosine similarity of these translations.

4 Experiments

We perform two sets of experiments, to investigate the effectiveness of our duality regularization in isolation (Section 4.2) and to compare our final models with the state-of-the-art methods in the literature (Section 4.3), respectively.

4.1 Experimental Settings

Dataset and Setup. Our datasets includes: (i) The Multilingual Unsupervised and Supervised Embeddings (**MUSE**) dataset released by Conneau et al. (2018). (ii) the more challenging **Vecmap** dataset from Dinu et al. (2015) and the extensions of Artetxe et al. (2017). We follow the evaluation setups of Conneau et al. (2018), utilizing cross-domain similarity local scaling (CSLS) for retrieving the translation of given source words. Following a standard evaluation practice (Vulić and Moens, 2013; Mikolov et al., 2013; Conneau et al., 2018), we report precision at 1 scores (P@1). Given the instability of existing methods, we follow Artetxe et al. (2018a) to perform 10 runs for each method and report the best and the average accuracies.

4.2 The Effectiveness of Dual Learning

We compare our method with Conneau et al. (2018) (Adv-C) under the same settings. As

¹We find that $\lambda = 0.5$ generally works well.

Setting	Adv-C		Ours		
	best	average.	best	average.	
MUSE	EN-ES	77.3	75.1	78.4	77.0
	ES-EN	79.1	73.5	79.0	75.6
	EN-DE	69.2	32.4	70.0	56.5
	DE-EN	68.5	31.7	69.3	53.7
	EN-IT	65.2	47.7	72.0	71.1
	IT-EN	64.0	45.3	69.9	69.4
	EN-EO	18.6	13.5	20.9	17.5
	EO-EN	16.6	12.0	17.3	15.3
Vecmap	EN-MS	17.9	08.3	24.7	21.8
	MS-EN	19.2	06.4	27.6	23.5
	EN-ES	26.2	20.5	29.6	26.1
	ES-EN	00.0	00.0	21.7	20.2
	EN-DE	40.3	20.0	43.7	36.5
	DE-EN	00.0	00.0	37.8	33.4
	EN-IT	38.3	37.0	38.5	37.5
	IT-EN	33.6	14.7	34.7	33.1
FI-EN	EN-FI	01.9	00.3	22.2	21.9
	FI-EN	00.0	00.0	20.0	18.9

Table 1: Accuracy on MUSE and Vecmap.

	EN-ES	EN-DE	EN-IT	EN-EO	EN-MS
Adv-C	66.95%	67.83%	70.23%	72.30%	75.87%
Ours	63.58%	64.29%	65.05%	64.06%	68.84%

Table 2: Inconsistency rates on MUSE.

Adv-C	Ours
three-tre-two	three-tre-three
neck-collo-ribcage	neck-collo-neck
door-finestrino-window	door-portiera-door
second-terzo-third	second-terzo-second
before-prima-first	before-dopo-after

Table 3: Word translation examples for English-Italian on MUSE. Ground truths are marked in **bold**.

shown in Table 1, our model outperforms Adv-C on both MUSE and Vecmap for all language pairs (except ES-EN). In addition, the proposed approach is less sensitive to initialization, and thus more stable than Adv-C over multiple runs. These results demonstrate the effectiveness of dual learning. Our method is also superior to Adv-C for the low-resource language pairs English \leftrightarrow Malay (MS) and English \leftrightarrow English-Esperanto (EO). Adv-C gives low performances on ES-EN, DE-EN, but much better results on the opposite directions on Vecmap. This is likely because the separate models are highly under-constrained, and thus easy to get stuck in poor local optima. In contrast, our method gives comparable results on both directions for the two languages, thanks to the use of information symmetry.

Table 2 shows the inconsistency rates² of

²For each word x_i from the source language, we check whether the primal \mathcal{F} and the dual mapping \mathcal{G} can recover x_i , i.e. $x_i \rightarrow \mathcal{F}(x_i) \rightarrow \mathcal{G}(\mathcal{F}(x_i)) \rightarrow x_i$.

Supervision	Approach	EN-IT		EN-DE		EN-FI		EN-ES	
		\rightarrow	\leftarrow	\rightarrow	\leftarrow	\rightarrow	\leftarrow	\rightarrow	\leftarrow
Supervised Methods	Procrustes	45.33	39.05	47.27	41.13	32.16	30.01	36.67	30.94
	GPA†	45.33	-	48.46	-	31.39	-	-	-
	GeoMM	48.17	41.10	49.40	44.73	36.03	38.24	39.27	34.58
	GeoMM _{semi}	50.00	42.67	51.47	46.96	36.24	39.57	39.30	36.06
Unsupervised Methods	Adv-C-Procrustes	45.40	38.78	46.40	00.00	25.21	00.15	35.47	0.05
	Unsup-SL	48.01	42.10	48.22	44.09	32.95	33.45	37.47	31.59
	Sinkhorn-BT	44.67	38.77	44.53	41.93	23.53	23.42	32.13	27.62
	Ours-Procrustes	45.60	38.29	46.58	42.50	28.08	26.48	35.20	28.94
	Ours-GeoMM _{semi}	50.00	42.67	51.60	47.22	35.88	39.62	39.47	36.43

Table 4: Accuracy (P@1) on **Vecmap**. The best results are **bolded**. †Results as reported in the original paper. For unsupervised methods, we report the average accuracy across 10 runs.

back translation between Adv-C and our method on MUSE. Compared with Adv-C, our model significantly reduces the inconsistency rates on all language pairs, which explains the overall improvement in Table 1. Table 3 gives several word translation examples. In the first three cases, our regularizer successfully fixes back translation errors. In the fourth case, ensuring cycle consistency does not lead to the correct translation, which explains some errors by our system. In the fifth case, our model finds a related word but not the same word in the back translation, due to the use of cosine similarity for regularization.

4.3 Comparison with the State-of-the-art

In this section, we compare our model with state-of-the-art systems, including those with different degrees of supervision. The baselines include: (1) **Procrustes** (Conneau et al., 2018), which learns a linear mapping through Procrustes Analysis (Schönemann, 1966). (2) **GPA** (Kementchedjhieva et al., 2018), an extension of Procrustes Analysis. (3) **GeoMM** (Jawanpuria et al., 2018), a geometric approach which learn a Mahalanobis metric to refine the notion of similarity. (4) **GeoMM_{semi}**, iterative GeoMM with weak supervision. (5) **Adv-C-Procrustes** (Conneau et al., 2018), which refines the mapping learned by Adv-C with iterative Procrustes, which learns the new mapping matrix by constructing a bilingual lexicon iteratively. (6) **Unsup-SL** (Artetxe et al., 2018a), which integrates a weak unsupervised mapping with a robust self-learning. (7) **Sinkhorn-BT** (Xu et al., 2018), which combines sinkhorn distance (Cuturi, 2013) and back-translation. For fair comparison, we integrate our model with two iterative refinement methods (Procrustes and GeoMM_{semi}).

Table 4 shows the final results on Vecmap.³ We first compare our model with the state-of-the-art unsupervised methods. Our model based on procrustes (Ours-Procrustes) outperforms Sinkhorn-BT on all test language pairs, and shows better performance than Adv-C-Procrustes on most language pairs. Adv-C-Procrustes gives very low precision on DE-EN, FI-EN and ES-EN, while Ours-Procrustes obtains reasonable results consistently. A possible explanation is that dual learning is helpful for providing good initiations, so that the procrustes solution is not likely to fall in poor local optima. The reason why Unsup-SL gives strong results on all language pairs is that it uses a robust self-learning framework, which contains several techniques to avoid poor local optima.

Additionally, we observe that our unsupervised method performs competitively and even better compared with strong supervised and semi-supervised approaches. Ours-Procrustes obtains comparable results with Procrustes on EN-IT and gives strong results on EN-DE, EN-FI, EN-ES and the opposite directions. Ours-GeoMM_{semi} obtains the state-of-the-art results on all tested language pairs except EN-FI, with the additional advantage of being fully unsupervised.

5 Conclusion

We investigated a regularization method to enhance unsupervised bilingual lexicon induction, by encouraging symmetry in lexical mapping between a pair of word embedding spaces. Results show that strengthening bi-directional mapping consistency significantly improves the effectiveness over the state-of-the-art method, leading to the best results on a standard benchmark.

³We select Vecmap as it is more challenging and closer to the real scenarios than MUSE (Artetxe et al., 2018a).

References

- Hanan Aldarmaki, Mahesh Mohan, and Mona Diab. 2018. *Unsupervised word mapping using structural similarities in monolingual embeddings*. *Transactions of the Association for Computational Linguistics*, 6:185–196.
- David Alvarez-Melis and Tommi Jaakkola. 2018. *Gromov-wasserstein alignment of word embedding spaces*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890. Association for Computational Linguistics.
- Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.
- 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. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. *A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. *Unsupervised statistical machine translation*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. *Unsupervised neural machine translation*. In *International Conference on Learning Representations*.
- A. P. Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *NIPS*.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. *Word translation without parallel data*. In *International Conference on Learning Representations*.
- Marco Cuturi. 2013. *Sinkhorn distances: Lightspeed computation of optimal transport*. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc.
- Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3th International Conference on Learning Representations (ICLR)*.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.
- Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. *Cross-lingual dependency parsing based on distributed representations*. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics.
- Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. *Dual learning for machine translation*. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc.
- Yedid Hoshen and Lior Wolf. 2018. *Non-adversarial unsupervised word translation*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478. Association for Computational Linguistics.
- Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2018. *Learning multilingual word embeddings in latent metric space: A geometric approach*. *CoRR*, abs/1808.08773.
- Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. *Generalizing procrustes analysis for better bilingual dictionary induction*. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220. Association for Computational Linguistics.
- Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. *Unsupervised machine translation using monolingual corpora only*. In *International Conference on Learning Representations*.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. *Phrase-based & neural unsupervised machine translation*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.

- Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. *Unsupervised cross-lingual information retrieval using monolingual data only*. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR ’18, pages 1253–1256, New York, NY, USA. ACM.
- Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. *Cheap translation for cross-lingual named entity recognition*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545. Association for Computational Linguistics.
- Antonio Valerio Miceli Barone. 2016. *Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders*. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126. Association for Computational Linguistics.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. *Exploiting similarities among languages for machine translation*. *CoRR*, abs/1309.4168.
- Tasnim Mohiuddin and Shafiq R. Joty. 2019. *Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training*. *CoRR*, abs/1904.04116.
- Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. *Learning unsupervised word translations without adversaries*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 627–632. Association for Computational Linguistics.
- Peter H. Schönemann. 1966. *A generalized solution of the orthogonal procrustes problem*. *Psychometrika*, 31(1):1–10.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Improving neural machine translation models with monolingual data*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96. Association for Computational Linguistics.
- Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116.
- Ivan Vulić and Marie-Francine Moens. 2015. *Mono-lingual and cross-lingual information retrieval models based on (bilingual) word embeddings*. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’15, pages 363–372, New York, NY, USA. ACM.
- Yijun Wang, Yingce Xia, Lu Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018. Dual transfer learning for neural machine translation with marginal distribution regularization. In *AAAI*.
- Yingce Xia, Jiang Bian, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. *Dual inference for machine learning*. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3112–3118.
- Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. *Neural cross-lingual named entity recognition with minimal resources*. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.
- Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. *Unsupervised cross-lingual transfer of word embedding spaces*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474. Association for Computational Linguistics.
- Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. *Unsupervised neural machine translation with weight sharing*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55. Association for Computational Linguistics.
- Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. *Adversarial training for unsupervised bilingual lexicon induction*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970. Association for Computational Linguistics.
- Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. *Earth mover’s distance minimization for unsupervised bilingual lexicon induction*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945. Association for Computational Linguistics.
- Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qi-Xing Huang, and Alexei A. Efros. 2016. Learning dense correspondence via 3d-guided cycle consistency. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 117–126.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.