MAAM: A Morphology-Aware Alignment Model for Unsupervised Bilingual Lexicon Induction

Pengcheng Yang[1,2], Fuli Luo[2], Peng Chen[2], Tianyu Liu[2], Xu Sun[1,2]
[1]Deep Learning Lab, Beijing Institute of Big Data Research, Peking University
[2]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
{yang_pc, luofuli, chen.peng, tianyu0421, xusun}@pku.edu.cn

Abstract

The task of unsupervised bilingual lexicon induction (UBLI) aims to induce word translations from monolingual corpora in two languages. Previous work has shown that morphological variation is an intractable challenge for the UBLI task, where the induced translation in failure case is usually morphologically related to the correct translation. To tackle this challenge, we propose a morphology-aware alignment model for the UBLI task. The proposed model aims to alleviate the adverse effect of morphological variation by introducing grammatical information learned by the pre-trained denoising language model. Results show that our approach can substantially outperform several state-of-the-art unsupervised systems, and even achieves competitive performance compared to supervised methods.

1 Introduction

The task of unsupervised bilingual lexicon induction aims at identifying translational equivalents across two languages (Kementchedjhieva et al., 2018). It can be applied in plenty of real-scenarios, such as machine translation (Artetxe et al., 2018b), transfer learning (Zhou et al., 2016), and so on.

Based on the observation that embedding spaces of different languages exhibit similar structures, a prominent approach is to align monolingual embedding spaces of two languages with a simple linear mapping (Zhang et al., 2017a; Lample et al., 2018). However, previous work (Artetxe et al., 2018a; Søgaard et al., 2018) has shown that morphological variation is an intractable challenge for the UBLI task. The induced translations in failure cases are usually morphologically related words. Due to similar semantics, these words can easily confuse the system to make the incorrect alignment. Table 1 presents three randomly selected failure examples of MUSE (Lample et al., 2018) on the FR-EN language pair, showing that all failures can be attributed to morphological variation. For instance, for the French source word "*mangez*", MUSE translates it to morphologically related word "*eats*", instead of the correct English translation "*eat*".

Source word	**Top-3 of retrieved nearest neighbors**		
mangez	eats	**eat**	buttered
suspendit	suspending	suspend	**suspended**
diffusant	broadcasts	broadcast	**broadcasting**

Table 1: Three randomly selected failure examples of MUSE on FR-EN language pair. **Red** words are correct translations, which are all not the nearest translations.

However, we find that additional grammatical information can help alleviate the adverse effect of morphological variation. In detail, since lexicon induction (word alignment) can be regarded as word-to-word translation, the fluency of the translated sentence can reflect the quality of word alignment. If the model can retrieve the correct translation for each word in a source sentence, the translated sentence is more likely to be fluent and grammatically correct. Considering some problems (e.g. word order) of the naive word-to-word translation can also lead to poor fluency, we pre-train a denoising auto-encoder (DAE) to clean noise in the original translated sentence. Figure 1 visually shows an example. For the French source word "*mangez*", if the model translates it to "*eats*" instead of the correct English translation "*eat*", the denoised translated sentence "*you eats meat*" is grammatically unreasonable. Therefore, by considering the fluency of the denoised translated sentence, these morphologically related erroneous translations can be reasonably punished.

Motivated by this, we propose a morphology-aware alignment model to alleviate the adverse effect of morphological variation by introducing ad-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3190–3196
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

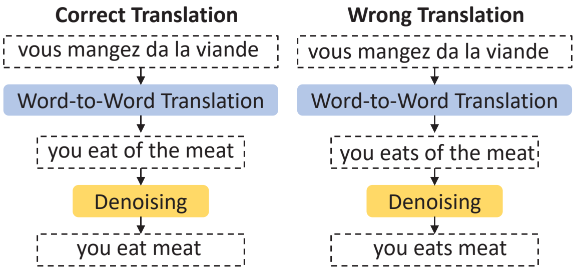

Figure 1: Alleviate the adverse effect of morphological variation via grammatical information.

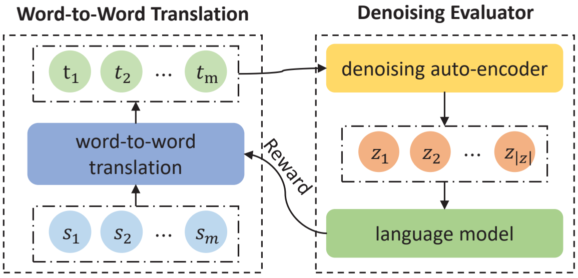

Figure 2: The sketch of the proposed model.

ditional grammatical information. The proposed model consists of a learnable linear transformation $\mathbf{W}$ between two languages and a parameter-fixed denoising evaluator $\mathbf{E}$. $\mathbf{W}$ is responsible for performing word-to-word translation on sentences in the source monolingual corpus. $\mathbf{E}$ first applies a DAE to clean noise in the original translated sentence, and then evaluates the fluency of the denoised translated sentence via a language model pre-trained on the target monolingual corpus to guide the training of $\mathbf{W}$. Due to the discrete operation of word-to-word translation, we employ REINFORCE algorithm (Williams, 1992) to estimate the corresponding gradient. With the grammatical information contained in $\mathbf{E}$, the adverse effect of morphological differences can be alleviated.

Our main contributions are listed as follows:

- We propose a morphology-aware alignment model for unsupervised bilingual lexicon induction, which aims to alleviate the adverse effect of morphological variation by introducing grammatical information learned from pre-trained language model.
- Extensive experimental results show that our approach achieves better performance than several state-of-the-art unsupervised systems, and even achieves competitive performance compared to supervised methods.

2 Proposed Model

We use $\mathcal{X} = \{x_i\}_{i=1}^{n_1}$ and $\mathcal{Y} = \{y_i\}_{i=1}^{n_2}$ to denote the source and target monolingual embeddings, respectively. The task aims to find a linear transformation $\mathbf{W}$ so that for any source word embedding x, $\mathbf{W}x$ lies close to the embedding y of its translation. Figure 2 presents the sketch of our proposed morphology-aware alignment model, which consists of a learnable linear transformation $\mathbf{W}$ and a parameter-fixed denoising evaluator $\mathbf{E}$.

2.1 Word-to-Word Translation

The word-to-word translation is accomplished by linear transformation $\mathbf{W}$. Specifically, for each word s_i in a source sentence $\boldsymbol{s} = (s_1, \cdots, s_m)$, it is translated by retrieving the nearest target word t_i based on cosine[1] similarity.

$$t_i = \operatorname*{argmax}_t \cos(\mathbf{W}x_{s_i}, y_t) \tag{1}$$

where x_{s_i} and y_t represent the pre-trained monolingual embedding of the source word s_i and target word t, respectively.

2.2 Denoising Evaluator

The denoising evaluator $\mathbf{E}$ aims to utilize learned grammar information to guild the training of $\mathbf{W}$. It contains two crucial components: a denoising auto-encoder (DAE) and a language model. Both components are pre-trained on the target monolingual corpus and remain fixed during training.

Denoising Auto-Encoder

Considering some ingrained problems (e.g. word order) of the naive word-to-word translation, the original translation $\boldsymbol{t}$ can be regarded as a noisy version of the ground-truth translation. Therefore, we adopt a DAE (Vincent et al., 2008) to clean noise in $\boldsymbol{t} = (t_1, \cdots, t_m)$ so that $\mathbf{E}$ can provide a more accurate supervisory signal. Here we implement the DAE as an encoder-decoder framework (Bahdanau et al., 2015). The input is the noisy version $\mathcal{N}(\boldsymbol{c})$ and the output is the cleaned sentence $\boldsymbol{c}$, where $\boldsymbol{c}$ is a sentence sampled from the target monolingual corpus. Following Kim et al. (2018), we construct $\mathcal{N}(\boldsymbol{c})$ by designing three noises: insertion, deletion, and reordering. Readers can refer to Kim et al. (2018) for more technical explanations.

[1] For simplicity, we employ the cosine similarity. Readers can also adopt other retrieval methods (e.g. CSLS) to obtain better performance.

Language Model

For a source sentence $\boldsymbol{s}$, if $\mathbf{W}$ is of high quality, the denoised translated sentence should keep fluent and grammatically correct. Otherwise, if $\mathbf{W}$ retrieves a morphologically related but erroneous word, the denoised translated sentence tends to be grammatically incorrect, leading to poor fluency. Therefore, a language model is used to evaluate the fluency of translation to guide the training of $\mathbf{W}$. We implement the language model as an LSTM (Hochreiter and Schmidhuber, 1997) structure with weight tying. Since this part is not the focus of our work, readers can refer to Press and Wolf (2017) for the details. With the grammatical information learned by the pre-trained language model, erroneous word alignment due to morphological variation is penalized. Therefore, $\mathbf{W}$ is encouraged to retrieve correct word translation with appropriate morphology.

2.3 Training and Testing

We encourage $\mathbf{W}$ to perform correct word alignment so that the denoised translated sentences are fluent and grammatically correct. Therefore, the training objective is to minimize the negative expected reward, which is formulated as follows:

$$\mathcal{L}(\boldsymbol{s}) = -\mathbb{E}_{\boldsymbol{t}}\Big[R(\boldsymbol{z_t})\text{log}p(\boldsymbol{t}|\boldsymbol{s})\Big] \qquad (2)$$

where $\boldsymbol{z_t}$ is the output of denoising auto-encoder with $\boldsymbol{t}$ as the input, $R(\boldsymbol{z_t})$ is the reward evaluating the fluency of $\boldsymbol{z_t}$, and $p(\boldsymbol{t}|\boldsymbol{s})$ is the probability of $\mathbf{W}$ outputs $\boldsymbol{t}$ by performing word-to-word translation on $\boldsymbol{s}$. We introduce them in detail as follows.

For the i-th word s_i in the source sentence $\boldsymbol{s}$, the probability of $\mathbf{W}$ retrieving the target translation t_i can be characterized by the cosine similarity of both embedding $\mathbf{W}x_{s_i}$ and y_{t_i}. Formally,

$$p(t_i|s_i) = \frac{\exp\big(\cos(\mathbf{W}x_{s_i}, y_{t_i})\big)}{\sum_t \exp\big(\cos(\mathbf{W}x_{s_i}, y_t)\big)} \qquad (3)$$

Therefore, $p(\boldsymbol{t}|\boldsymbol{s})$ can be defined as the product of the probability corresponding to each position:

$$p(\boldsymbol{t}|\boldsymbol{s}) = \prod_{i=1}^{m} p(t_i|s_i) \qquad (4)$$

The reward $R(\boldsymbol{z_t})$ aims at evaluating the fluency of the denoised translated sentence $\boldsymbol{z_t}$ to guide the training of $\mathbf{W}$, which is defined as follows:

$$R(\boldsymbol{z_t}) = \exp\Big(\frac{1}{|\boldsymbol{z_t}|}\sum_{i=1}^{|\boldsymbol{z_t}|}\text{log}q(z_i|\boldsymbol{z}_{<i})\Big) \qquad (5)$$

where z_i is the i-th word in $\boldsymbol{z_t} = (z_1, \cdots, z_{|\boldsymbol{z}|})$, $\boldsymbol{z}_{<i}$ refers to the sequence $(z_1, \cdots, z_{i-1})$, and $q(z_i|\boldsymbol{z}_{<i})$ is the probability that the pre-trained language model outputs the word z_i conditioned on $\boldsymbol{z}_{<i}$. If $\boldsymbol{z_t}$ is fluent and grammatically correct, the corresponding reward $R(\boldsymbol{z_t})$ is relatively large. Therefore, the reward $R(\boldsymbol{z_t})$ can be used as feedback to guide the training of $\mathbf{W}$. Since operation of word-to-word translation is discrete, we use REINFORCE algorithm (Williams, 1992) to estimate the gradient of Eq. (2) as follows:

$$\nabla_{\mathbf{W}}\mathcal{L}(\boldsymbol{s}) \approx -\big(R(\boldsymbol{z_t}) - b\big)\nabla_{\mathbf{W}}\text{log}\big(p(\boldsymbol{t}|\boldsymbol{s})\big) \qquad (6)$$

where b is the baseline that is responsible for reducing the variance of gradient estimate.

3 Experiments

3.1 Experiment Settings

We conduct experiments on the 300-dim *fastText* embeddings trained on Wikipedia. All words are lower-cased and only the frequent 200K words are used. We utilize approach in Artetxe et al. (2018a) to provide the initial linear transformation and lexicon constructed by Lample et al. (2018) is used for evaluation. Here we report accuracy with *nearest neighbor retrieval* based on cosine similarity. The parameters of the DAE and language model are provided in the Appendix. We set the batch size to 64 and the optimizer is SGD. The learning rate is initialized to 10^{-5} and it is halved after every training epoch. The unsupervised criterion proposed in Lample et al. (2018) is adopted as both a stopping criterion and a model selection criterion.

3.2 Experimental Results

Table 2 presents the results of different systems, showing that our proposed model achieves the best performance on all test language pairs under unsupervised settings. In addition, our approach is able to achieve completely comparable or even better performance than supervised systems. This illustrates that the quality of word alignment can be improved by introducing grammar information from the pre-trained denoising language model. Our denoising evaluator encourages the model to retrieve the correct translation with appropriate morphological by assessing the fluency of sentences obtained by word-to-word translation. This alleviates the adverse effect of morphological variation.

Methods	DE-EN	EN-DE	ES-EN	EN-ES	FR-EN	EN-FR	IT-EN	EN-IT
Supervised:								
Mikolov et al. (2013a)	61.93	**73.07**	74.00	**80.73**	71.33	**82.20**	68.93	**77.60**
Xing et al. (2015)	67.73	69.53	77.20	78.60	76.33	78.67	72.00	73.33
Shigeto et al. (2015)	**71.07**	63.73	**81.07**	74.53	**79.93**	73.13	**76.47**	68.13
Artetxe et al. (2016)	69.13	72.13	78.27	80.07	77.73	79.20	73.60	74.47
Artetxe et al. (2017)	68.07	69.20	75.60	78.20	74.47	77.67	70.53	71.67
Unsupervised:								
Zhang et al. (2017a)	40.13	41.27	58.80	60.93	-	57.60	43.60	44.53
Zhang et al. (2017b)	-	55.20	70.87	71.40	-	-	64.87	65.27
Lample et al. (2018)	69.73	71.33	79.07	78.80	77.87	78.13	74.47	75.33
Xu et al. (2018)	67.00	69.33	77.80	79.53	75.47	77.93	72.60	73.47
Artetxe et al. (2018a)	72.27	73.60	81.60	80.67	80.20	80.40	76.33	77.13
Ours	**73.13**	**74.47**	**82.13**	**81.87**	**81.53**	**81.27**	**77.60**	**78.33**

Table 2: The accuracy of different methods in various language pairs. **Bold** indicates the best supervised and unsupervised results, respectively. "-" means that the model fails to converge and hence the result is omitted.

Models	EN-ES	EN-FR	EN-DE	EN-IT
Full model	81.87	81.27	74.47	78.33
w/o Evaluator	80.67	80.40	73.60	77.13
w/o DAE	81.33	80.93	74.20	77.73

Table 3: Results of ablation study.

3.3 Ablation Study

Here we perform an ablation study to understand the importance of different components. Table 3 presents the performance of different ablated versions, showing that our denoising evaluator can bring stable improvements in performance. This illustrates that introducing grammatical information learned by the pre-trained denoising language model is of great help to perform accurate word alignment. By imposing the penalty to the retrieved morphologically related but erroneous translations, this additional grammatical information can alleviate the adverse effects of morphological variation. In addition, we can find that the DAE plays an active role in improving results. By cleaning the noise in the original translated sentence, the DAE makes the reward provided by evaluator more accurate, leading to the improvements in model performance.

3.4 The Validity of Cleaning Noise

By cleaning the noise in the original word-to-word translation, the denoising auto-encoder (DAE) can benefit the evaluator **E** to feedback more accurate evaluation signals. Here Table 4 presents several examples output by the DAE on the FR-EN language pair. The results show that there exist some obvious grammatical errors in the naive word-to-word translation. For instance, the word "*to*" is missing in the first example and the words in the second example are not organized in a grammatical order. However, our pre-trained DAE is able to correct these errors by inserting or deleting appropriate words or adjusting the word order. This intuitively demonstrates the effectiveness of our DAE in cleaning noise contained in the original translated sentence.

Input: *Être adulte, c'est être seul.*
Noisy translation: *Be adult, it's be alone.*
Cleaned translation: *To be an adult is to be alone.*
Ground truth: *To be an adult is to be alone.*

Input: *L'histoire se répète.*
Noisy translation: *History itself repeats.*
Cleaned translation: *History repeats itself.*
Ground truth: *History repeats itself.*

Table 4: Several examples output by the denoising auto-encoder on the FR-EN language pair.

3.5 Case Study

Table 5 lists several word translation examples on the FR-EN language pair. The results show that the baselines retrieve morphologically related but erroneous translations, while our approach is able to perform the correct word alignment. Our approach can constrain the retrieved translation to have the correct morphology by introducing grammatical information, leading to improved performance. Figure 3 presents the visualization of joint semantic space of FR-EN language pair using t-SNE (Maaten and Hinton, 2008), showing that word pairs that can be translated mutually are represented by almost the same point. This intuitively reveals that our approach can capture the common linguistic regularities of different languages.

Source word	MUSE	Vecmap	Ours
suspendit	suspending	suspend	**suspended**
diffusant	broadcasts	broadcast	**broadcasting**
atteint	reaching	reach	**reached**

Table 5: Translations of various systems on the FR-EN language pair. **Red** words are correct translations.

4 Related Work

This paper is mainly related to the following two lines of work.

Supervised cross-lingual embedding. Inspired by the isometric observation between monolingual word embeddings of two different languages, Mikolov et al. (2013b) propose to learn cross-lingual word mapping by minimizing mean squared error. Latter, Dinu and Baroni (2015) investigate the hubness problem and Faruqui and Dyer (2014) incorporates the semantics of a word in multiple languages into its embedding. Furthermore, Xing et al. (2015) propose to impose the orthogonal constraint to the linear mapping and Artetxe et al. (2016) present a series of techniques, including length normalization and mean centering, to improve bilingual results. There also exist some other representative researches. For instance, Smith et al. (2017) present inverse-softmax which normalizes the softmax probability over source words rather than target words and Artetxe et al. (2017) present a self-learning framework to perform iterative refinement, which is also adopted in some unsupervised settings and plays a crucial role in improving performance.

Unsupervised cross-lingual embedding. The endeavors to explore unsupervised cross-lingual embedding are mainly divided into two categories. One line focuses on designing heuristics or utilizing the structural similarity of monolingual embeddings. For instance, Hoshen and Wolf (2018) present a non-adversarial method based on the principal component analysis. Both Aldarmaki et al. (2018) and Artetxe et al. (2018a) take advantage of geometric properties across languages to perform word retrieval to learn the initial word mapping. Cao and Zhao (2018) formulate this problem as point set registration to adopt a point set registration method. However, these methods usually require plenty of random restarts or additional skills to achieve satisfactory performance. Another line strives to learn unsupervised word mapping by direct distribution-matching. For example, Lample et al. (2018) and Zhang et al. (2017a) completely eliminate the need for any supervision signal by aligning the distribution of transferred embedding and target embedding with GAN. Furthermore, Zhang et al. (2017b) and Xu et al. (2018) adopt the Earth Mover's distance and Sinkhorn distance as the optimized distance metrics, respectively. There are also some attempts on distant language pairs. For instance, Kementchedjhieva et al. (2018) generalize Procrustes analysis by projecting the two languages into a latent space and Nakashole (2018) propose to learn neighborhood sensitive mapping by training non-linear functions. As for the hubness problem, Ruder et al. (2018) propose a latent-variable model learned with Viterbi EM algorithm. Recently, Alaux et al. (2018) work on the problem of aligning more than two languages simultaneously by a formulation ensuring composable mappings.

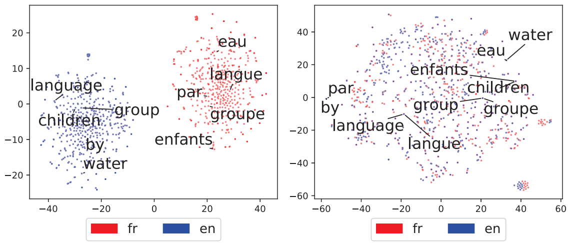

Figure 3: Visualization of two monolingual embedding spaces (**left**) and aligned embedding space (**right**).

5 Conclusion

In this work, we present a morphology-aware alignment model for unsupervised bilingual lexicon induction. The proposed model is able to alleviate the adverse effect of morphological variation by introducing grammatical information learned from pre-trained denoising language model. The results show that our approach can achieve better performance than several state-of-the-art unsupervised systems, and even achieves competitive performance compared to supervised methods.

Acknowledgement

We thank the anonymous reviewers for their thoughtful comments. We also would like to thank Shuangzhi Wu and Dongdong Zhang for their insightful suggestions. Xu Sun is the contact author of this paper.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyperalignment for multilingual word embeddings. *arXiv preprint arXiv:1811.01124*.

Hanan Aldarmaki, Mahesh Mohan, and Mona T. Diab. 2018. Unsupervised word mapping using structural similarities in monolingual embeddings. *TACL*, 6:185–196.

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

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. 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.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *6th International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

Hailong Cao and Tiejun Zhao. 2018. Point set registration for unsupervised bilingual lexicon induction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3991–3997.

Georgiana Dinu and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *3rd International Conference on Learning Representations*.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yedid Hoshen and Lior Wolf. 2018. An iterative closest point method for unsupervised word translation. *arXiv preprint arXiv:1801.06126v1*.

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.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 862–868.

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*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Ndapa Nakashole. 2018. NORMA: neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers*, pages 157–163.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 458–468.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference*, pages 135–151.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulic. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 778–788.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference*, pages 1096–1103.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

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

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.

Pengcheng Yang, Fuli Luo, Shuangzhi Wu, Jingjing Xu, Dongdong Zhang, and Xu Sun. 2018. Learning unsupervised word mapping by maximizing mean discrepancy. *arXiv preprint arXiv:1811.00275*.

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.

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.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. 2018. Learning sentiment memories for sentiment modification without parallel data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1108.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*.