ATOGAN:Adaptive Training Objective Generative Adversarial Network for Cross-lingual Word Alignment in Non-Isomorphic Embedding Spaces

Anonymous ACL submission

Abstract

Cross-lingual word alignment is a task for word translation from monolingual word embedding spaces of two languages. Recent works are mostly based on supervised approaches, which need specific bilingual seed dictionaries. The unsupervised adversarial approaches, which utilize the generative adversarial networks to map the whole monolingual space, do not need any aligned data. However these approaches pay no attention to the problem of mode collapse and gradient disappearance in generative adversarial networks(GAN). We proposed an adaptive training objective generative adversarial network(ATOGAN). We combined particle swarm optimization(PSO) with GAN to select the training objective in GAN's training, which alleviates the problem of mode collapse and gradient disappearance. Moreover, we improved the word alignment by bi-directional mapping and consistency loss. Experimental results demonstrate that our approach is better than several state-of-the-art approaches in distant language pairs(non-isomorphic embedding spaces).

1 Introduction

Currently, cross-lingual word alignment plays an important role in language understanding and generation tasks for various Natural Language Processing (NLP) applications, such as cross-lingual named entity recognition, cross-lingual sentiment analysis and cross-lingual text classification, etc. Cross-lingual word alignment transfers the embedding spaces between language pairs to address resource lack in monolingual corpora. This task in non-isomorphic embedding spaces is a great challenge at present.

In isomorphic assumption, monolingual corpora have the similar structures across languages by the training of word embedding (Mikolov et al., 2013), and then different monolingual word embedding spaces could be transformed with each other. However, those of etymologically and typologically distant languages are far from isomorphic.

Supervised approaches are commonly used for the cross-lingual word alignment in the early works. The monolingual space is aligned by bilingual seed dictionaries((Vulić et al., 2019), (Glavaš et al., 2019), (Glavaš and Vulić, 2020)). Compared with supervised learning approaches, unsupervised learning approaches for the cross-lingual word alignment do not need specific bilingual dictionaries or parallel corpora. They utilized generative adversarial networks(GAN) (Goodfellow et al., 2014) to address this task, which have achieved great in the deep learning field. GAN is a combination of two neural networks, the generator and discriminator, which are trained against each other to generate realistic synthetic real-valued data. Based on this work, Barone (2016) first proposed an unsupervised adversarial method in cross-lingual word alignment task. In Zhang et al. (2017), they improved the training of GAN by changing the linear transformation matrix to an orthogonal matrix and adding noise to the word vector. This work allows GAN to be better trained. In Lample et al. (2018), they proposed a refined approach after the training of GAN and cross-domain similarity local scaling(CSLS) for word translation retrieval, it improved the results a lot. In Bai et al. (2019), they transformed the source and the target monolingual word embeddings into a shared embedding space. In Li et al. (2021), they proposed a noise function to disperse dense word embeddings and a Wasserstein critic network to preserve the semantics of the source word embeddings. However, these methods did not consider that the isomorphic assumption might not always hold, especially those of etymologically and typologically distant languages, which are far from isomorphic.(Glavaš and Vulić, 2020). What's more, in the training of

GAN, there have always been problems of mode collapse and gradient disappearance. This problem is more obvious in distant languages task.

In order to address mode collapse and gradient disappearance in GAN's training, researchers have done a lot of improvement by developing various adversarial training objectives. LSGAN (Mao et al., 2017) presents least squares loss function to overcome the problems of vanishing gradients. Wasserstein GAN(Arjovsky et al., 2017) and Wasserstein GAN gradient penalty (WGAN-GP) (Gulrajani et al., 2017) used Wasserstein distance and gradient-penalty in adversarial training goals. However, they all use a predefined single adversarial objective function. Evolutionary GAN (Wang et al., 2019) combined evolution strategy with GAN for the first time. It utilizes different training objectives as mutation operations to jointly optimize the generator for improving both the training stability and generative performance. Based on evolutionary GAN, CatGAN (Liu et al., 2020) utilize a hierarchical evolutionary learning algorithm for training the model and obtaining the balance between the sample quality and diversity. Multi-Objective Evolutionary Generative Adversarial Network (MO-EGAN)(Baioletti et al., 2020) re-defined the evaluation of generators as a multi-objective problem to address the conflict of quality and diversity. However, this methods only utilized a little evolutionary strategy. The training objective of the generator is a simple choice, and it is hard to choose a better combination of training objective.

This paper proposes an adaptive training objective generative adversarial network(ATOGAN) for cross-lingual word alignment in non-isomorphic embedding spaces. We find the non-consistency of mapping from source to target and target to source can effect the accuracy of word alignment. For instance, the word 'horse' maps to 'Chevaux' in French, while 'Chevaux' maps to 'sheep' in English. Thus, we utilize the cycle consistency loss in Cycle-GAN (Zhu et al., 2017) and improve it. The generator is used as a mapper to map the embedding space from source language to the target language. The discriminator distinguishes whether the embedding is from the source or the target language. This model can reduce the cycle consistency loss which represent the non-consistency in language pairs. To address the mode collapse and gradient disappearance in the training for the task of distant langauge pairs, we design a dynamic loss function for the training objective by particle swarm optimization(PSO) which is an evolutionary computation. The training objective is a combination of various adversarial loss and cycle consistency loss. In various epoch of training, we will use the PSO algorithm to calculate an optimal parameter combination of loss weights to adjust the training objective.

The contributions of this paper are as follows:

- We propose a novel unsupervised cross-lingual word alignment approach through consistency loss and bi-directional mapping in non-isomorphic embedding spaces.
- We improve the generate network via particle swarm optimization to select the better training objective in various training periods, which alleviates mode collapse and gradient disappearance in GAN's training.
- Extensive experiments are performed on dataset. Compared with previous adversarial approaches in distant language pairs tasks, the proposed approach is more effective.

2 Methodology

In our approach, the source language word embedding is set to $x \in X$, while the target is $y \in Y$. The proposed model is a bi-directional GAN with evolutionary computing, as Figure 1. The bi-GAN consists of discriminator_A, discriminator_B, generator(G) and generator(F). The procedure of evolutionary computing by PSO comprises variation, evaluation and selection step.

2.1 bi-directional mapping with consistency loss

The generator(G) and generator(F) are both the orthogonal matrix which performance better in cross-lingual word alignment task. G maps x to y, while F maps y to x. Discriminator_A and discriminator_B are trained against generators by judging that the word embedding is real or not. The word alignment from source to target is the opposite direction task of the alignment from target to source. The better results of one direction mapping can improved the opposite mapping results.

$$L_{cycle} = E_{x \to X}[||F(G(x)) - x||] + E_{y \to Y}[||G(F(y)) - y||] \quad (1)$$

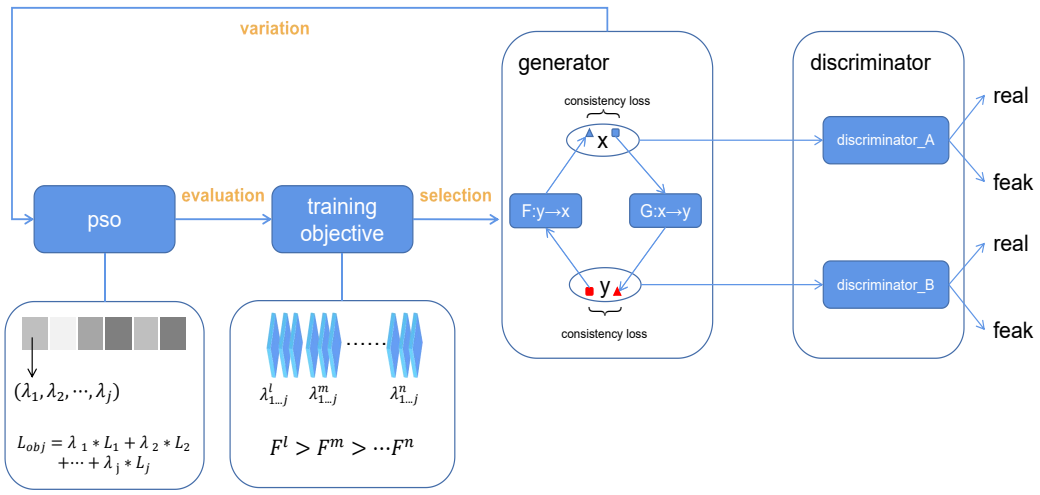

Figure 1: Architecture of unsupervised cross-lingual word alignment approach based on cycle-GAN and hybrid training.

$$L_{id} = E_{x\to X}[||F(x) - x||] + E_{y\to Y}[||G(y) - y||] \quad (2)$$

Inspired by cycle-GAN, we utilize consistency loss to constraint the consistency of bi-directional mapping. Adversarial training can learn the mapping between two monolingual space. However, with large enough embedding space capacity, although the word embedding space has been aligned as a whole, some words are only aligned to the same target word. This words are not accurately aligned to the translated words. If one word is mapped to its near-synonym in target language, which is not its accurate translated word, the two-way direction mapping would cause non-consistency. Thus, reducing consistency loss can alleviate this problem, the loss is showed in formula (2). In addition, we hope that the mapping outputs belong to target space for any inputs, so we use identity loss to assist mapping, as formula (3).

$$L_{gen1}(X, G, D\_B) = -E_{x\to X}[log(D\_B(G(x)))] \quad (3)$$

$$L_{gen2}(X, Y, G, D\_B) = -E_{x,y\to X,Y}[D\_B(y)log(D\_B(G(x))) + (1 - D\_B(y)log(1 - D\_B(G(x))))] \quad (4)$$

$$L_{gen3}(X, G, D\_B) = -E_{x\to X}[(D\_B(G(x)) - 1)^2] \quad (5)$$

$$L_{obj} = \lambda_1 * L_{gen1} + \lambda_2 * L_{gen2} + \lambda_3 * L_{gen3} + \lambda_4 * L_{cycle} + \lambda_5 * L_{id} \quad (6)$$

$$L_D(X, Y, G, D\_B) = E_{x\to X}[log(D\_B(G(x)))] + E_{y\to Y}[log(1 - D\_B(y))] \quad (7)$$

2.2 adaptive training objective via PSO

Most existing GANs address the mode collapse and gradient disappearance in GAN's training via transforming the adversarial objective function. Inspired by EGAN, different objective function will achieve the best results in various training periods. The single adversarial objective function could cause mode collapse or gradient disappearance in a certain training period. We design a combination of

multiple different loss functions as training objective, then we utilize the PSO algorithm to select various weight combination of loss functions in various training periods. As Eq. (6), we combine three various adversarial loss(Eq. (3)(4)(5)) and cycle consistency loss(Eq. (1)), identity loss(Eq. (2)) as the final training objective function(L_{obj}). It should be noted that the EGAN only select one loss function as objective in various training periods. It equals to the weight combination of one value being 1, the other being 0 in our approach. The weight combination of loss functions include the selecting of only one adversarial loss function.

$$F(X, G, D\_B) = E_{x \to X}[(D\_B(G(x))] - \lambda_d * log||\nabla_{D\_B}|| \quad (8)$$

We utilize a fitness function to score the weight combination in each particle training, then save the best result when the training of PSO is ending. The fitness function is combined of the quality of generated data and generative diversity, as Eq. (8). For evaluating the quality, the prediction of generated data is higher, the quality is better. It measures the gap between the generated samples and the real samples. For evaluating the diversity, the minus log-gradient-norm of optimizing D is utilized to measure the diversity of generated samples. Corresponding to a small discriminator gradient, the generated data tend to be scattered enough to avoid obvious countermeasures for the discriminator.

In the training, we utilize the average cosine similarity in language pairs evaluate the results of each epoch. If the value of average cosine similarity less than maximum value k times continuously, we consider the objective is not the better in current training period, then the PSO will start to find the best objective function. Thus, the training objective is adaptive in the GAN's training by the selecting in PSO.

3 Experiment

3.1 Experimental Setup

Data. We evaluate on the MUSE dataset introduced by Lample et al. (2018). The MUSE dataset consists of monolingual word embeddings of dimension 300 trained with fasttext on Wikipedia corpora and gold dictionaries for 110 language pairs. According to BLISS(Patra et al., 2019), the high Gromov-Hausdorff distance language pairs(distant language pairs) cannot be aligned well using orthogonal transforms. We choose a part of distant language pairs such as: English (En) from/to Russian(ru), Danish(da), Indonesian(id), Hungary(hu) and Croatia(hr).

Baselines. We compare ATOGAN to the state-of-the-art unsupervised cross-lingual word alignment approaches:(1)Adv-C(Lample et al., 2018) proposed CSLS for selecting the nearest neighbor of vector, meanwhile it utilize procrustes analysis to refine the learned mapping by the dictionary which is build from unsupervised model. (2)Adv-B(Bai et al., 2019) trains two separate auto-encoders to map two monolingual embeddings into a shared embedding space. (3)Adv-L(Li et al., 2021) introduces a noise function that can disperse the dense embedding points and utilize a Wasserstein critic network to encourage adding noise.

Implementation Details and Hyperparameter Tuning. We consider the most frequent 200k word embeddings for evaluation. We take cross-domain similarity local scaling as the retrieval metrics and use the average cosine similarity between these deemed translations as a validation metric. The number of particle is set to 20 and the iterations of PSO is set to 50. ATOGAN has two hyperparameter: the threshold K for activating PSO, the weight of diversity in fitness function λ_d. We verified that setting K = 2 and λ_d = 0.4 performs best for most language pairs.

3.2 Results and Discussion

We report the results of average word translation Precision@1 on Table 1. We compare our approach with state-of-the-art unsupervised approaches. The results show that the proposed approach outperforms other unsupervised approaches on 9 of 10 mapping directions in distant language pairs. It can be seen that the two direction mapping results in the same language pair have a big difference, for instance, the average word translation Precision@1 of en-hu is 54.7% while hu-en is 66.0%. It means a certain one direction mapping is harder than other in the same language pair. We consider the bi-directional mapping can use one direction mapping to assist training other direction mapping in the same language pair.

model	en-ru	ru-en	en-da	da-en	en-hu	hu-en	en-hr	hr-en	en-id	id-en
Adv-C	44.0	59.1	58.1	51.9	53.5	62.9	*	*	67	62.2
Adv-B	49.0	**65.8**	57.7	64.6	52.5	63.0	33.4	48.2	68.0	68.5
Adv-L	29.9	43.2	-	-	47.1	62.4	32.5	48.0	-	-
ATOGAN	**49.9**	63.4	**59.3**	**66.7**	**54.7**	**66.0**	**35.9**	**48.7**	**68.6**	**68.7**

Table 1: Average word translation precision retrieval P@1(%) on MUSE. Best results are bolded.(The best results of 5 runs,'*' denotes an precision of less than 0.1%. '-' denotes that we can not get the results.)

model	id-en		da-en		hu-id	
	P@1	s	P@1	s	P@1	s
Adv-C	62.3	1	11.2	1	61.3	6
ATOGAN	**68.0**	**9**	**66.7**	**10**	**65.5**	**9**

Table 2: The average word translation precision retrieval P@1 (%) and the number of successful runs (those with >5% accuracy) on MUSE. Best results are bolded.(The average results of 10 runs)

λ_d	en-ru	ru-en	en-id	id-en
0.1	49.3	62.3	67.8	68.5
0.2	48.5	61	68.1	68.3
0.4	**49.9**	62.9	**68.6**	**68.7**
0.8	48.8	**63.4**	68.1	68.4

Table 3: Average word translation precision retrieval P@1(%) on MUSE. Best results are bolded.(The best results of 5 runs)

In order to verify the stable of our approach, we report the average result of the word translation Precision@1 and the number of successful runs((those with >5% accuracy)) in 10 runs on Table 2. The results show that our approach can successfully run 9 or 10 times in 10 runs for id-en,da-en and hu-en, while on the same language pairs and mapping direction, the Adv-C can only run a small number of times successfully, what's more, our average Precision@1 is also higher, which is close to best value in Table 1. We consider that the adaptive training objective let the GAN's training more stable by alleviating mode collapse and gradient disappearance in GAN's training.

To further analyze our model, we perform ablation studies and show the results for two language pairs(en from/to ru, en from/to id) on Table 3 and Figure 2. Table 3 shows the performance for different λ_d, which is a weight for fitness function in PSO. We need a balance of quality and diversity when using PSO to select the weight of objective function. The λ_d is used to adjust it. The results show that when the λ_d is set to 0.4, we can get the best results on most of experiments. However, due to the diversity of different languages, one value of this hyperparameter is hard to fit all language pairs.

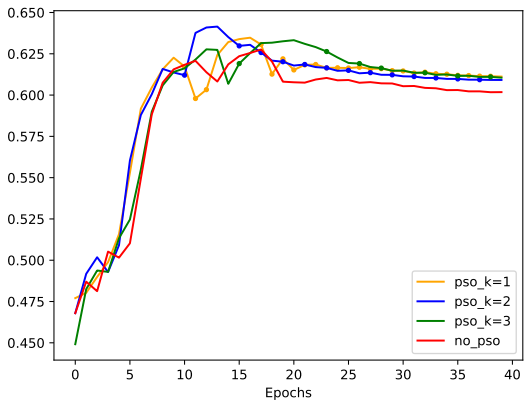

Figure 2: The average cosine similarity in each epochs for en to ru. We add the point where PSO is activated.

We perform the results of different PSO threshold K and without PSO for en to ru in Figure 2. The results show that when the weight of training objective is improved by PSO, the decreasing trend can be improved a lot. Compared to the results of without PSO, adaptive training objective can reach to a better value. We consider that the various loss function have different priority in various training periods. Using PSO to select the weight of training objective can help the model find a better training direction.

4 Conclusion

In this paper, we propose an adaptive training objective generative adversarial network for cross-lingual word alignment in non-isomorphic embedding spaces. Our approach uses consistency loss and bi-directional mapping to improve the orthogonal mapping. We combined PSO with GAN to alleviate mode collapse and gradient disappearance in GAN's training, so that the mapping is stable. The experiments show that our approach performs

better than other strong baselines in distant language pairs(non-isomorphic). In the future we will focus on the combined of supervised approaches and our approach to improve the mapping on cross-lingual word alignment tasks.

References

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *ICML*.

Xuefeng Bai, Hailong Cao, Kehai Chen, and Tiejun Zhao. 2019. A bilingual adversarial autoencoder for unsupervised bilingual lexicon induction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1639–1648.

Marco Baioletti, Carlos Artemio Coello Coello, Gabriele Di Bari, and Valentina Poggioni. 2020. Multi-objective evolutionary gan. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, pages 1824–1831.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *ACL 2016*, page 121.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *NIPS*.

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

Yuling Li, Yuhong Zhang, Kui Yu, and Xuegang Hu. 2021. Adversarial training with wasserstein distance for learning cross-lingual word embeddings. *Applied Intelligence*, pages 1–13.

Zhiyue Liu, Jiahai Wang, and Zhiwei Liang. 2020. Catgan: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8425–8432.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

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

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418.

Chaoyue Wang, Chang Xu, Xin Yao, and Dacheng Tao. 2019. Evolutionary generative adversarial networks. *IEEE Transactions on Evolutionary Computation*, 23(6):921–934.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. 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.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.