WSD-GAN: Word Sense Disambiguation Using Generative Adversarial Networks

Zijian Hu,[1,2]* Fuli Luo,[1]* Yutong Tan,[2] Wenxin Zeng,[3] Zhifang Sui[1]
[1]Key Laboratory of Computational Linguistics, Ministry of Education,
School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2]Beijing Normal University, Beijing, China
[3]Beijing University of Posts and Telecommunications
zijian_hu@mail.bnu.edu.cn, luofuli@pku.edu.cn, tanyt@mail.bnu.edu.cn,
zengwenxin@bupt.edu.cn, szf@pku.edu.cn

Abstract

Word Sense Disambiguation (WSD), as a tough task in Natural Language Processing (NLP), aims to identify the correct sense of an ambiguous word in a given context. There are two mainstreams in WSD. Supervised methods mainly utilize labeled context to train a classifier which generates the right probability distribution of word senses. Meanwhile knowledge-based (unsupervised) methods which focus on glosses (word sense definitions) always calculate the similarity of context-gloss pair as score to find out the right word sense. In this paper, we propose a generative adversarial framework WSD-GAN which combines two mainstream methods in WSD. The generative model, based on supervised methods, tries to generate a probability distribution over the word senses. Meanwhile the discriminative model, based on knowledge-based methods, focuses on predicting the relevancy of the context-gloss pairs and identifies the correct pairs over the others. Furthermore, in order to optimize both two models, we leverage policy gradient to enhance the performances of the two models mutually. Our experimental results show that WSD-GAN achieves competitive results on several English all-words WSD datasets.

Introduction

There are several research lines about WSD. We can divide them into two categories: **knowledge-driven** unsupervised methods and **data-driven** supervised methods.

- Knowledge, especially the gloss which actually defines a word sense meaning, proves helpful to WSD. The gloss-based methods like Lesk (Basile, Caputo, and Semeraro 2014) solves the WSD problem by picking the highest overlap (or similarity) between the context and gloss.
- Labeled data plays a key role in supervised methods which treat WSD as a classification task(Kågebäck and Salomonsson 2016). These models are trained beyond traditional log-likelihood and predict a word sense by choosing the highest probability over the word sense distribution.

In this paper, we consider the two schools of thinking as two sides of the same coin. Inspired by Generative Adversarial Nets (GANs) using in Information Retrieval (Wang et al. 2017), we propose a novel model **WSD-GAN** which unifies the above mentioned two methods. Unlike the recent works which incorporate knowledge into neural network (Luo et al. 2018b), we adopt a *minimax game* theory. The generator aims to maximise the log-likelihood of classification by learning from labelled data, while the discriminator acts as a challenger which takes advantage of knowledge information and further pushes the generator to its limit. Experimental results show that the WSD-GAN takes the advantage of both the two kinds of resources and then better combine the two schools of thinking in WSD.

*Equal Contribution.

WSD-GAN

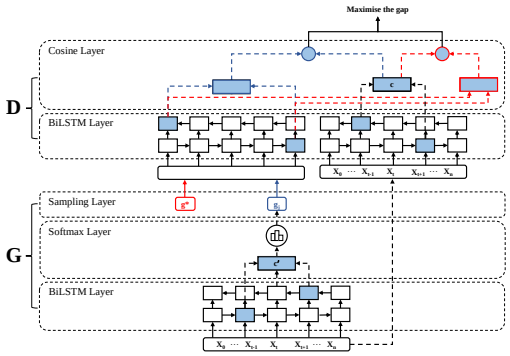

Figure 1: Architecture of WSD-GAN

Architecture of WSD-GAN

The WSD-GAN architecture is illustrated in Fig 1, which contains two separated models, the generator and the discriminator, represented as G and D.

Generator The input $[x_0, x_1, ..., x_t, ..., x_{n-1}, x_n]$ is the context words around the target ambiguous words. Having utilized the pre-trained embedding matrix, we feed them into the bidirectional LSTM Layer. The representation of the context c is computed as the concatenation of the forward output at position $t-1$ and the backward output at position $t+1$. More specifically, $c = [\overrightarrow{h}_{t-1} : \overleftarrow{h}_{t+1}]$.

And then, we leverage a softmax layer to compute the probability $p_\theta(g|c)$ digstribution over all the senses (or

glosses g).

$$p_\theta(g|c) = softmax(Wc + b)$$

In the Sampling Layer, we sample m specific senses according to the probability distribution and feed the glosses of the sampled senses into the discriminator.

Discriminator According to the m sampled word senses from the generator, the corresponding glosses are also encoded by LSTM. The gloss representation vector g is concatenated by last units of both forward and backward LSTM outputs. The context representation vector c is computed the same way in the generator. Then, the score $f_\phi(g, c)$ of the each context and gloss pair is computed as the cosine similarity of g and c.

$$f_\phi(g, c) = \cos\langle g, c\rangle$$

Optimization Objective

Overall Objective For the generator, we want to generate a sample distribution to minimize the cross-entropy loss of classification. For the discriminator, we want to maximize the gap between the correct sense and the fake sense of the ambiguous word. Therefore, the total objective function is optimized as:

$$J^{G*,D*} = \min_\theta \max_\phi \sum_{n=1}^{N} (E_{g\sim p_true(g|c_n,r)}[\log D(g|c_n)] + E_{g\sim p_\theta(g|c_n,r)}[\log(1 - D(g|c_n))])$$

where $D(g|c)$ is computed as

$$D(g|c) = \sigma(f_\phi(g, c)) = \frac{\exp(f_\phi(g, c))}{1 + \exp(f_\phi(g, c))}$$

Discriminator Optimization The purpose of objective function is to maximize the log-likelihood of correctly distinguishing the true and generated gloss. Therefore, the objective function for discriminator is:

$$\phi* = \arg\max_\phi \sum_{\mathrm{n=1}}^{\mathrm{N}} (E_{g\sim p_{true}(g|c_n,r)}[\log D(g|c_n)] + E_{g\sim p_\theta(g|c_n,r)}[\log(1 - D(g|c_n))])$$

which can be optimized by mini-batch gradient descent.

Generator Optimization The purpose of objective function is to minimize the objective function.

$$\begin{aligned}\theta* &= \arg\min_\theta \sum_{n=1}^{N} \left(E_{g\sim\ p_{true}(g|c_n,r)}[\log D(g|c_n)] + E_{d\sim\ p_\theta(g|c_n,r)}[\log(1 - D(g|c_n))]\right) \\ &= \arg\max_\theta \sum_{n=1}^{N} E_{g\sim p\theta(g|c_n,r)}[\log(1 + \exp(f_\phi(g, c_n)))]\end{aligned}$$

Experiments

Dataset We train our model on the SemCor 3.0 which manually annotated based on the WordNet 3.0. And we do evaluation on several English all-words WSD datasets which include Senseval-2 (SE2), Senseval-3 (SE3), SemEval-07 (SE7), SemEval-13 (SE13), and SemEval-15 (SE15).

System	SE2	SE3	SE13	SE15	All
MFS baseline	65.6	66.0	63.8	67.1	65.5
Lesk$_{+emb}$	63.0	63.7	66.2	64.6	64.2
BiLSTM	71.1	68.4	64.8	68.3	68.4
GAS	72.1	70.2	67.0	71.8	70.3
GAS$_{ext}$	72.2	70.5	67.2	72.6	**70.6**
WSD-GAN	73.3	70.2	66.5	72.0	**70.6**

Table 1: F1-score (%) for English all-words WSD.

Analysis The Table 1 shows that WSD-GAN achieves the overall result 70.6 on the four test datasets, which is competitive to the state-of-the-art WSD system GAS$_{ext}$ (Luo et al. 2018b). However, we use less gloss knowledge than GAS$_{ext}$. For a fair comparison, our WSD-GAN can beat GAS which use the same gloss as us. Furthermore, compared to the best knowledge-based methods Lesk$_{+emb}$ (Basile, Caputo, and Semeraro 2014), our model improves the overall result by 6.4%. Compared to the supervised neural-based methods BiLSTM (Kågebäck and Salomonsson 2016), our model improves the overall result by 2.2%. It proves that our the integration of GAN and WSD can achieve great improvement.

Conclusions

In this paper, we combine the knowledge-driven unsupervised methods and data-driven supervised methods of WSD into a unified framework via GAN.

Acknowledgments

This paper is supported by NSFC project 61751201 and M1752013. The contact author is Zhifang Sui.

References

Basile, P.; Caputo, A.; and Semeraro, G. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Roceedings of COLING 2014, the International Conference on Computational Linguistics: Technical Papers*.

Kågebäck, M., and Salomonsson, H. 2016. Word sense disambiguation using a bidirectional lstm. *arXiv preprint arXiv:1606.03568*.

Luo, F.; Liu, T.; He, Z.; Xia, Q.; Sui, Z.; and Chang, B. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1402–1411.

Luo, F.; Liu, T.; Xia, Q.; Chang, B.; and Sui, Z. 2018b. Incorporating glosses into neural word sense disambiguation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Wang, J.; Yu, L.; Zhang, W.; Gong, Y.; Xu, Y.; Wang, B.; Zhang, P.; and Zhang, D. 2017. IRGAN: A minimax game for unifying generative and discriminative information retrieval models. *CoRR* abs/1705.10513.