Adversarial Discriminative Denoising for Distant Supervision Relation Extraction

Bing Liu[1], Huan Gao[1], Guilin Qi[1,2(✉)], Shangfu Duan[1], Tianxing Wu[3], and Meng Wang[1]

[1] School of Computer Science and Engineering, Southeast University, Nanjing 211111, China
{liubing_cs,hg,gqi,sf_duan,meng.wang}@seu.edu.cn
[2] Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 211111, China
[3] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore
wutianxing@ntu.edu.sg

Abstract. Distant supervision has been widely used to generate labeled data automatically for relation extraction by aligning knowledge base with text. However, it introduces much noise, which can severely impact the performance of relation extraction. Recent studies have attempted to remove the noise explicitly from the generated data but they suffer from (1) the lack of an effective way of introducing explicit supervision to the denoising process and (2) the difficulty of optimization caused by the sampling action in denoising result evaluation. To solve these issues, we propose an adversarial discriminative denoising framework, which provides an effective way of introducing human supervision and exploiting it along with the potentially useful information underlying the noisy data in a unified framework. Besides, we employ a continuous approximation of sampling action to guarantee the holistic denoising framework to be differentiable. Experimental results show that very little human supervision is sufficient for our approach to outperform the state-of-the-art methods significantly.

Keywords: Distant supervision · Relation extraction · Noise reduction · Adversarial discriminative model

1 Introduction

Distant supervision (DS) is a promising approach to relation extraction (RE). It can generate training data automatically by aligning knowledge base (KB) with text [3,6]. However, it suffers from the introduced noise, which can severely effect the performance of RE. Previous researches have focused on building DS models with noise adaptability [2,7]. However, they did not remove the noise explicitly. To enable explicit noise reduction, a few challenges need to be addressed.

G. Li et al. (Eds.): DASFAA 2019, LNCS 11448, pp. 282–286, 2019.
https://doi.org/10.1007/978-3-030-18590-9_29

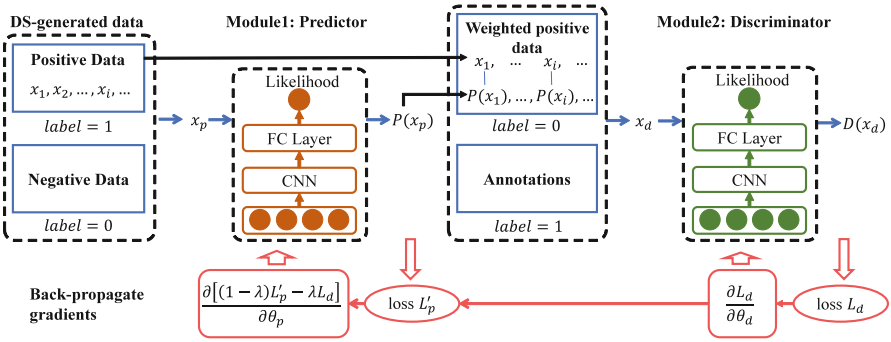


Fig. 1. Overview of the adversarial discriminative denoising framework.

The first challenge comes from the lack of an effective way of introducing explicit supervision to the denoising process. Without introducting human supervision, the unsupervised methods can only make a coarse-grained distinction between the true positive instances and the noise. Another challenge relates to the evaluation of the denoising result. The existing approaches [1,4,5] performed the evaluation by sampling the noisy data according to the noise recognizer and then assessing the resulting subset. The problem is that the sampling action can lead to non-differentiability, which hinders the use of the evaluators that back-propagate gradients to guide the optimization of the noise recognizer in a holistic manner.

To solve the above challenges, we propose an adversarial discriminative denoising framework, which can not only acquire the denoising ability by exploiting the beneficial information underlying DS-generated data but also further get boosted via introducing very few human annotations efficiently. To guarantee the model to be differentiable, we employ a continuous approximation of sampling action when evaluating the denoising result, which helps fast convergence and gains a better solution.

2 Adversarial Discriminative Denoising Framework

The overview of our approach is shown in Fig. 1. The DS-generated dataset to be cleansed can be partitioned into positive data $\mathcal{D}^p$ and negative data $\mathcal{D}^n$ concerning a specific relation r. The adversarial learning process is carried out on a handful of manually labeled data $\mathcal{D}^l$ and a mass of DS-generated data. The framework consists of two core modules: (1) a true positive instance predictor P, which acts as a noise recognizer and outputs the probability that an instance is true positive, and (2) a data source discriminator D, which functions as a critic of the distinguishability between $\mathcal{D}^l$ and $\mathcal{D}^p$ weighted by P. P not only tries to satisfy the supervision from DS but also attempts to assign $\mathcal{D}^p$ with proper weights to make it indistinguishable from $\mathcal{D}^l$. D attempts to improve its ability to distinguish $\mathcal{D}^l$ from the weighted $\mathcal{D}^p$ and back feed the similarity of

these two data sets to P. When evaluating the denoising result of P, D assigns weights to the instances in $\mathcal{D}^p$ instead of sampling $\mathcal{D}^p$ according to its outputting probabilities.

2.1 Predictor

To obtain a valid P, we train it using two sources of guidance: (1) the DS-generated labeled data (this supervision information is actually from KB) and (2) the feedback from D which represents the distinguishability between $\mathcal{D}^l$ and the weighted $\mathcal{D}^p$. Although the DS-generated data contain much noise, they can provide beneficial information due to the correctly labeled instances. To take advantage of this information, we use P to classify $\mathcal{D}^p$ and $\mathcal{D}^n$ and take the classification loss as part of its loss function. Another goal of P is to reduce the distinguishability between $\mathcal{D}^l$ and the weighted $\mathcal{D}^p$. Thus, we treat this distinguishability, which is measured by D, as another part of the loss of P.

2.2 Continuous Approximation of Sampling Action

We approximate the sampling action by assigning each instance x in $\mathcal{D}^p$ with $P(x)$ as the weight and let the weights play a role in measuring the similarity between $\mathcal{D}^p$ and $\mathcal{D}^l$. In this continuous approximation setting, the instances with higher weights have more effect on the measurement, which is equivalent to more frequent participation in sampling setting. Therefore, this similarity is controlled by the weights.

2.3 Discriminator

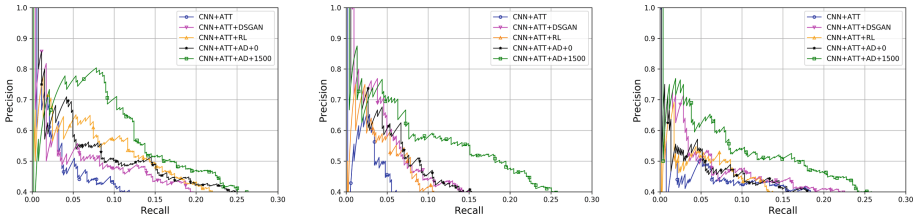
D aims to detect if an instance is from $\mathcal{D}^l$ or the weighted $\mathcal{D}^p$. Essentially, D is the metrics of the weights, and P adjusts itself according to the feedback about the weights. In $\mathcal{D}^p$, the true positive instances are more difficult to be correctly recognized by D than the false positive ones. In order to puzzle D and cause more losses to it, P will assign higher weights on the true positive ones while lower weights on the noisy ones. As the adversary of P, D has to pay more attention to the instances with high weights so as to avoid major losses. This will drive P to avert mistaking noisy data as the correct ones.

2.4 Cleaning Noisy Dataset with Predictor

After the adversarial learning process, we obtain a valid P concerning relation r. Then, we apply P to $\mathcal{D}^p$ and filter out the instances whose scores are below a certain threshold thr. The cleansed positive data will be used as the positive data of relation r in the training stage of RE models.

3 Experiments

Our experiments are carried out on the widely used NYT dataset[1] [6]. Following the previous work [3], we evaluated our method on NYT using the held-out evaluation. We chose a CNN based model with attention mechanism (CNN+ATT) proposed by [2] as the DS RE model and used this CNN+ATT model trained on the original NYT dataset as the baseline. Then, we chose two state-of-the-art denoising methods based on RL [5] (CNN+ATT+RL) and GAN [4] (CNN+ATT+DSGAN) for comparison. To show the effect of human annotations, we trained our approach in two ways, including using none annotation (CNN+ATT+AD+0) and 1500 annotations (CNN+ATT+AD+1500). Figure 2 shows the PR curves of held-out evaluation. It demonstrates that our approach can effectively reinforce the existing models and outperforms state-of-the-art methods.



(a) 31% noise proportion (b) 50% noise proportion (c) 70% noise proportion

Fig. 2. Aggregate PR curves of CNN+ATT based model upon the held-out evaluation. Our approach (CNN+ATT+AD+1500) outperforms state-of-the-art methods significantly.

4 Conclusion

In this paper, we proposed an adversarial discriminative denoising framework to remove the false positive noise explicitly from DS-generated data. This framework provides an effective way of introducing human supervision and exploiting it along with the potentially useful information underlying the original data in a unified framework.

Acknowledgement. This work was supported by National Key R&D Program of China (2018YFC0830200) and National Natural Science Foundation of China Key Project (U1736204).

[1] http://iesl.cs.umass.edu/riedel/ecml/.

References

1. Feng, J., Huang, M., Zhao, L., Yang, Y., Zhu, X.: Reinforcement learning for relation classification from noisy data. In: Proceedings of AAAI, pp. 5779–5786 (2018)
2. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: Proceedings of ACL, pp. 2124–2133 (2016)
3. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of ACL, pp. 1003–1011 (2009)
4. Qin, P., Xu, W., Wang, W.Y.: DSGAN: generative adversarial training for distant supervision relation extraction. In: Proceedings of ACL, pp. 496–505 (2018)
5. Qin, P., Xu, W., Wang, W.Y.: Robust distant supervision relation extraction via deep reinforcement learning. In: Proceedings of ACL, pp. 2137–2147 (2018)
6. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
7. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of EMNLP, pp. 1753–1762 (2015)