Active Testing: An Unbiased Evaluation Method for Distantly Supervised Relation Extraction

Pengshuai Li[1], **Xinsong Zhang**[2], **Weijia Jia**[3,1*] and **Wei Zhao**[4]
[1]Dept. of CSE, Shanghai Jiao Tong University, Shanghai, China [2]ByteDance AI Lab
[3]Institute of AI & Future Networks, Beijing Normal University (Zhuhai) & UIC, PR China
[4]American University of Sharjah, Sharjah, United Arab Emirates
pengshuai.li@sjtu.edu.cn zhangxinsong.0320@bytedance.com
jiawj@bnu.edu.cn wzhao@aus.edu

Abstract

Distant supervision has been a widely used method for neural relation extraction for its convenience of automatically labeling datasets. However, existing works on distantly supervised relation extraction suffer from the low quality of test set, which leads to considerable biased performance evaluation. These biases not only result in unfair evaluations but also mislead the optimization of neural relation extraction. To mitigate this problem, we propose a novel evaluation method named active testing through utilizing both the noisy test set and a few manual annotations. Experiments on a widely used benchmark show that our proposed approach can yield approximately unbiased evaluations for distantly supervised relation extractors.

1 Introduction

Relation extraction aims to identify relations between a pair of entities in a sentence. It has been thoroughly researched by supervised methods with hand-labeled data. To break the bottleneck of manual labeling, distant supervision (Mintz et al., 2009) automatically labels raw text with knowledge bases. It assumes that if a pair of entities have a known relation in a knowledge base, all sentences with these two entities may express the same relation. Clearly, the automatically labeled datasets in distant supervision contain amounts of sentences with wrong relation labels. However, previous works only focus on wrongly labeled instances in training sets but neglect those in test sets. Most of them estimate their performance with the held-out evaluation on noisy test sets, which will yield inaccurate evaluations of existing models and seriously mislead the model optimization. As shown in Table 1, we compare the results of held-out evaluation and human evaluation for the same model on a widely used benchmark dataset NYT-10 (Riedel et al., 2010). The biases between human evaluation and existing held-out evaluation are over 10%, which are mainly caused by wrongly labeled instances in the test set, especially false negative instances.

Evaluations	P@100	P@200	P@300
Held-out Evaluation	83	77	69
Human Evaluation	93(+10)	92.5(+15.5)	91(+22)

Table 1: The Precision at top K predictions (%) of the model Lin et al. (2016) upon held-out evaluation and human evaluation on NYT-10. Results are obtained by our implementations.

A false negative instance is an entity pair labeled as non-relation, even if it has at least one relation in reality. This problem is caused by the incompleteness of existing knowledge bases. For example, over 70% of people included in Freebase have no place of birth (Dong et al., 2014). From a random sampling, we deduce that about 8.75% entity pairs in the test set of NYT-10 are misclassified as non-relation.[1] Clearly, these mislabeled entity pairs yield biased evaluations and lead to inappropriate optimization for distantly supervised relation extraction.

In this paper, we propose an active testing approach to estimate the performance of distantly supervised relation extraction. Active testing has been proved effective in evaluating vision models with large-scale noisy datasets (Nguyen et al., 2018). In our approach, we design an iterative approach, with two stage per iteration: vetting stage and estimating stage. In the vetting stage, we adopt an active strategy to select batches of the most valuable entity pairs from the noisy test set for annotating. In the estimating stage, a metric estimator is proposed to obtain a more accurate evaluation.

*Corresponding author: jiawj@bnu.edu.cn.

[1]We randomly selected 400 entity pairs from the test set, in which 35 are misclassified as non-relation.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 204–211
November 16 - 20, 2020.

With a few vetting-estimating iterations, evaluation results can be dramatically close to that of human evaluation by using limited vetted data and all noisy data. Experimental results demonstrate that the proposed evaluation method yields approximately unbiased estimations for distantly supervised relation extraction.

2 Related Work

Distant supervision (Mintz et al., 2009) was proposed to deal with large-scale relation extraction with automatic annotations. A series of studies have been conducted with human-designed features in distantly supervised relation extraction (Riedel et al., 2010; Surdeanu et al., 2012; Takamatsu et al., 2012; Angeli et al., 2014; Han and Sun, 2016). In recent years, neural models were widely used to extract semantic meanings accurately without hand-designed features (Zeng et al., 2015; Lin et al., 2017; Zhang et al., 2019). Then, to alleviate the influence of wrongly labeled instances in distant supervision, those neural relation extractors integrated techniques such as attention mechanism (Lin et al., 2016; Han et al., 2018; Huang and Du, 2019), generative adversarial nets (Qin et al., 2018a; Li et al., 2019), and reinforcement learning (Feng et al., 2018; Qin et al., 2018b). However, none of the above methods pay attention to the biased and inaccurate test set. Though human evaluation can yield accurate evaluation results (Zeng et al., 2015; Alt et al., 2019), labeling all the instances in the test set is too costly.

3 Task Definition

In distant supervision paradigm, all sentences containing the same entity pair constitute a bag. Researchers train a relation extractor based on bags of sentences and then use it to predict relations of entity pairs. Suppose that a distantly supervised model returns confident score[2] $s_i = \{s_{i1}, s_{i2} \dots s_{ip}\}$ for entity pair $i \in \{1 \dots N\}$, where p is the number of relations, N is the number of entity pairs, and $s_{ij} \in (0, 1)$. $y_i = \{y_{i1}, y_{i2} \dots y_{ip}\}$ and $z_i = \{z_{i1}, z_{i2} \dots z_{ip}\}$ respectively represent automatic labels and true labels for entity pair i, where y_{ij} and z_{ij} are both in $\{0, 1\}$[3].

In widely used held-out evaluation, existing methods observe two key metrics which are precision at top K ($P@K$) and Precision-Recall curve (*PR curve*). To compute both metrics, confident score for all entity pairs are sorted in descending order, which is defined as $s' = \{s'_1, s'_2 \dots s'_P\}$ where $P = Np$. Automatic labels and true labels are denoted as $y' = \{y'_1, \dots, y'_P\}$ and $z' = \{z'_1, \dots, z'_P\}$. In summary, $P@K$ and $R@K$ can be described by the following equations,

$$P@K\{z'_1 \dots z'_P\} = \frac{1}{K} \sum_{i \le K} z'_i \quad (1)$$

$$R@K\{z'_1 \dots z'_P\} = \frac{\sum_{i \le K} z'_i}{\sum_{i \le P} z'_i} \quad (2)$$

Held-out evaluation replaces z' with y' to calculate $P@K$ and $R@K$, which leads to incorrect results obviously.

4 Methodology

In this section, we present the general framework of our method. A small random sampled set is vetted in the initial state. In each iteration there are two steps: 1) select a batch of entity pairs with a customized vetting strategy, label them manually, and add them to the vetted set; 2) use a new metric estimator to evaluate existing models by the noisy set and the vetted set jointly. After a few vetting-evaluating iterations, unbiased performance of relation extraction is appropriately evaluated. In summary, our method consists of two key components: a vetting strategy and a metric estimator.

4.1 Metric Estimator

Our test set consists of two parts: 1) a noisy set U in which we only know automatic label y'_i; 2) a vetted set V in which we know both automatic label y'_i and manual label $\tilde{z}'_i$. We treat the true label z'_i as a latent variable and $\tilde{z}'_i$ is its observed value. The performance evaluation mainly depends on the estimation of z'_i. In our work, we estimate the probability as

$$p(z'_i) = \prod_{i \in U} p(z'_i|\Theta) \prod_{i \in V} \delta(z'_i = \tilde{z}'_i) \quad (3)$$

where Θ represents all available elements such as confident score, noisy labels and so on. We make the assumption that the distribution of true latent labels is conditioned on Θ.

Given posterior estimates $p(z'_i|\Theta)$, we can compute the expected performance by replacing the true

[2]Confident scores are estimated probabilities for relations.
[3]An entity pair may have more than one relations.

latent label by its probability. Then, the precision and recall equations can be rewritten as

$$E[P@K] = \frac{1}{K}(\sum_{i \in V_K} \tilde{z}'_i + \sum_{i \in U_K} p(z'_i = 1|\Theta)) \quad (4)$$

$$E[R@K] = \frac{\sum_{i \in V_K} \tilde{z}'_i + \sum_{i \in U_K} p(z'_i = 1|\Theta)}{\sum_{i \in V} \tilde{z}'_i + \sum_{i \in U} p(z'_i = 1|\Theta)} \quad (5)$$

where U_K and V_K denote the unvetted and vetted subsets of K highest-scoring examples in the total set $U \cup V$.

To predict the true latent label z'_i for a specific relation, we use noisy label y'_i and confident score s'_i. This posterior probability can be derived as (see appendix for proof)

$$p(z'_i|y'_i, s'_i) = \frac{p(y_{jk}|z_{jk})p(z_{jk}|s_{jk})}{\sum_v p(y_{jk}|z_{jk} = v)p(z_{jk} = v|s_{jk})} \quad (6)$$

where $v \in \{0, 1\}$. s_{jk}, y_{jk}, z_{jk} are the corresponding elements of s'_i, y'_i, z'_i before sorting confident score. Given a few vetted data, we fit $p(y_{jk}|z_{jk})$ by standard maximum likelihood estimation (counting frequencies). $p(z_{jk}|s_{jk})$ is fitted by using logistic regression. For each relation, there is a specific logistic regression function to fit.

4.2 Vetting Strategy

In this work, we apply a strategy based on *maximum expected model change*(MEMC) (Settles, 2009). The vetting strategy is to select the sample which can yield a largest expected change of performance estimation. Let $E_{p(z'|V)}Q$ be the expected performance based on the distribution $p(z'|V)$ estimated from current vetted set V. After vetting example i and updating that estimator, it will become $E_{p(z'|V,z'_i)}Q$. The change caused by vetting example i can be written as

$$\Delta_i(z'_i) = |E_{p(z'|V)}Q - E_{p(z'|V,z'_i)}Q| \quad (7)$$

For precision at top K, this expected change can be written as

$$E_{p(z'_i|V)}[\Delta_i(z'_i)] = \frac{2}{K} p_i(1 - p_i) \quad (8)$$

where $p_i = P(z'_i = 1|\Theta)$. For the PR curve, every point depends on $P@K$ for different K. Thus, this vetting strategy is also useful for the PR curve.

With this vetting strategy, the most valuable data is always selected first. Therefore, vetting budget is the only factor controlling the vetting procedure. In this approach, we take it as a hyper parameter. When the budget is used up, the vetting stops. The procedure is described in Algorithm 1.

Algorithm 1 Active Testing Algorithm

Require: unvetted set U, vetted set V, vetting budget T, vetting strategy *VS*, confident score S, estimator $p(z')$

1: **while** $T > 0$ **do**
2: select a batch of items $B \in U$ with vetting strategy *VS*
3: vet B and get manual label $\tilde{z}'$
4: U=U−B, V=V∪B
5: fit $p(z')$ with U, V, S
6: T=T−$|B|$
7: **end while**

5 Experiment

We conduct sufficient experiments to support our claims; 1) The proposed active testing is able to get more accurate results by introducing very few manual annotations. 2) The held-out evaluation will misdirect the optimization of relation extraction, which can be further proved through re-evaluation of eight up-to-date relation extractors.

5.1 Experimental Setting

Dataset. Our experiments are conducted on a widely used benchmark NYT-10 (Riedel et al., 2010) and an accurate dataset named NYT-19, which contains 500 randomly selected entity pairs from the test set of NYT-10. It contains 106 positive entity pairs and 394 negative entity pairs, in which 35 entity pairs are false negative. NYT-19 has been well labeled by NLP researchers.

Initialization. We use PCNN+ATT (Lin et al., 2016) as baseline relation extractors. To be more convincing, we provide the experimental results of BGRU+ATT in the appendix. The initial state of vetted set includes all the positive entity pairs of the test set in NYT-10 and 150 vetted negative entity pairs. The batch size for vetting is 20 and the vetting budget is set to 100 entity pairs.

5.2 Effect of Active Testing

We evaluate the performance of PCNN+ATT with held-out evaluation, human evaluation and our method. The results are shown in Table 2, and Figure 1. Due to high costs of manual labeling for

the whole test set, we use the PR-curve on NYT-19 to simulate that on NYT-10.

Model	Evaluations	P@100	P@200	P@300
PCNN+ATT	Held-out Evaluation	83	77	69
	Our method	91.2	88.4	83.4
	Human Evaluation	93	92.5	91

Table 2: The Precision at top K predictions (%) of PCNN+ATT upon held-out evaluation, our method and human evaluation on NYT-10.

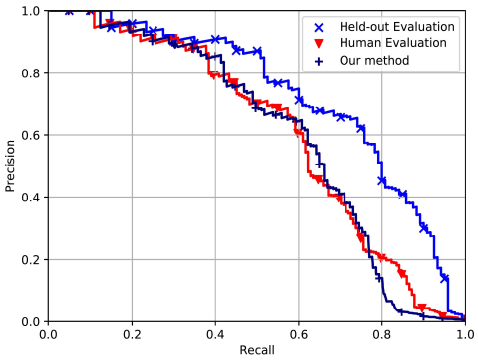

Figure 1: The PR curve of PCNN+ATT on NYT-19.

To measure the distance between two curves, we sample 20 points equidistant on each curve and calculate the Euclidean distance of the two vectors. In this way, our method gets the distances 0.17 to the curve of human evaluation while corresponding distances for held-out evaluation is 0.72. We can observe that 1) The performance biases between manual evaluation and held-out evaluation are too significant to be neglected. 2) The huge biases caused by wrongly labeled instances are dramatically alleviated by our method. Our method obtains at least 8.2% closer precision to manual evaluation than the held-out evaluation.

5.3 Effect of Vetting Strategy

We compare our MEMC strategy with a random vetting strategy as shown in Figure 2. The distance from curves of different vetting strategies to that of human evaluation is 0.176 and 0.284. From the figure, we can conclude that the proposed vetting strategy is much more effective than the random vetting strategy. With the same vetting budget, MEMC gets more accurate performance estimation at most parts of the range.

5.4 Re-evaluation of Relation Extractors

With the proposed performance estimator, we re-evaluate eight up-to-date distantly supervised relation extractors.

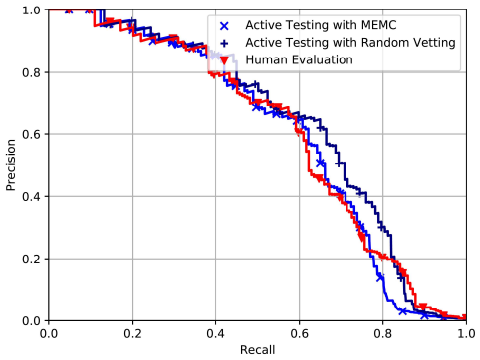

Figure 2: The PR curves of PCNN+ATT evaluated with various vetting strategies on NYT-19

Model	P@100(%)	P@200(%)	P@300(%)
Zeng et al. 2015	88.0	85.1	82.3
Lin et al. 2016	91.2	88.9	83.8
Liu et al. 2017	94.0	89.0	87.0
Qin et al. 2018b	88.8	86.2	84.8
Qin et al. 2018a	87.0	83.8	80.8
Liu et al. 2018	**95.7**	**93.4**	**89.9**
BGRU	94.4	89.5	84.7
BGRU+ATT	95.1	90.1	87.1

Table 3: The P@N precision of distantly supervised relation extractors on NYT-10. All the methods are implemented with the same framework and running in the same run-time environment.

From Table 3, we can observe that: 1) The relative ranking of the models according to precision at top K almost remains the same except Qin et al. 2018b and Qin et al. 2018a. Although GAN and reinforcement learning are helpful to select valuable training instances, they are tendentiously to be overfitted. 2) Most models make the improvements as they mentioned within papers at high confident score interval. 3) BGRU performs better than any other models, while BGRU based method Liu et al. 2018 achieves highest precision. More results and discussions can be found in the Appendix.

6 Conclusion

In this paper, we propose a novel active testing approach for distantly supervised relation extraction, which evaluates performance of relation extractors with both noisy data and a few vetted data. Our experiments show that the proposed evaluation method is appropriately unbiased and significant for optimization of distantly relation extraction in future.

Acknowledgements

This work is partially supported by Chinese National Research Fund (NSFC) Key Project No. 61532013 and No. 61872239; BNU-UIC Institute of Artificial Intelligence and Future Networks funded by Beijing Normal University (Zhuhai) and AI and Data Science Hub, BNU-HKBU United International College (UIC), Zhuhai, Guangdong, China.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398.

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1556–1567.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 601–610.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 5779–5786.

Xianpei Han and Le Sun. 2016. Global distant supervision for relation extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2950–2956.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2236–2245.

Yuyun Huang and Jinhua Du. 2019. Self-attention enhanced cnns and collaborative curriculum learning for distantly supervised relation extraction. 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 389–398.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Hai Zhao. 2019. Gan driven semi-distant supervision for relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3026–3035.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 34–43.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2124–2133.

Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. 2018. Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2204.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1790–1795.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 1003–1011.

Phuc Xuan Nguyen, Deva Ramanan, and Charless C. Fowlkes. 2018. Active testing: An efficient and robust framework for estimating accuracy. In *ICML*, pages 3759–3768.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. Dsgan: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–505.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2137–2147.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 148–163.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 455–465.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 721–729.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1753–1762.

Xinsong Zhang, Pengshuai Li, Weijia Jia, and Hai Zhao. 2019. Multi-labeled relation extraction with attentive capsule network. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 3243–3249.

A Appendices

A.1 Logistic Regression

Here we provide the derivation of Equation.6 in the main paper.

$$
\begin{aligned}
p(z_i'|y_i', s_i') &= \frac{p(z_i', y_i', s_i')}{\sum_v p(z_i' = v, y_i', s_i')} \\
&= \frac{p(z_{jk}, y_{jk}, s_{jk})}{\sum_v p(z_{jk} = v, y_{jk}, s_{jk})} \\
&= \frac{p(y_{jk}|z_{jk}, s_{jk})p(z_{jk}|s_{jk})}{\sum_v p(y_{jk}|z_{jk} = v, s_{jk})p(z_{jk} = v|s_{jk})}
\end{aligned}
$$

We assume that given z_{jk}, the observed label y_{jk} is conditionally independent of s_{jk}, which means $p(y_{jk}|z_{jk}, s_{jk}) = p(y_{jk}|z_{jk})$. The expression is simplified to:

$$
p(z_i'|y_i', s_k') = \frac{p(y_{jk}|z_{jk})p(z_{jk}|s_{jk})}{\sum_v p(y_{jk}|z_{jk} = v)p(z_{jk} = v|s_{jk})}
$$

A.2 Vetting Strategy

Here we provide the derivation of Equation.8 in the main paper.

$$
\begin{aligned}
E_{p(z_i'|V)}[\Delta_i(z_i')] &= p_i \frac{1}{K}|1 - p_i| + (1 - p_i)\frac{1}{K}|0 - p_i| \\
&= \frac{2}{K}p_i(1 - p_i)
\end{aligned}
$$

Model	Evaluations	P@100	P@200	P@300
BGRU+ATT	Held-out Evaluation	82	78.5	74.3
	Our method	95.2	90.1	87.1
	Human Evaluation	98	96	95

Table 4: The Precision at top K predictions (%) of BGRU+ATT upon held-out evaluation, our method and human evaluation on NYT-10.

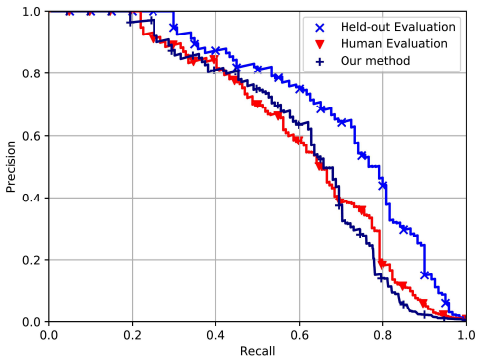

Figure 3: The PR curve of BGRU+ATT on NYT-19.

A.3 Experimental result of BGRU+ATT

We also evaluate the performance of BGRU+ATT with held-out evaluation, human evaluation and our method. The results are shown in Table 4, and Figure 3. Our method gets the distances 0.15 to the curve of human evaluation while corresponding distances for held-out evaluation is 0.55.

A.4 The result of different iterations

We have recorded the distance of different iterations between the curves obtained by our method and manual evaluation in Figure 4. With the results, we can observe that the evaluation results obtained by our method become closer to human evaluation when the number of annotated entity pairs is less than 100. When the number is more than 100, the distance no longer drops rapidly but begins to fluctuate.

B Case Study

We present realistic cases in NYT-10 to show the effectiveness of our method. In Figure 6, all cases are selected from Top 300 predictions of PCNN+ATT. These instances are all negative instances and has the automatic label NA in NYT-10. In held-out evaluation, relation predictions for these instances are judged as wrong. However, part of them are false negative instances in fact and have the corresponding relations, which cause considerable biases between manual and held-out evaluation. In

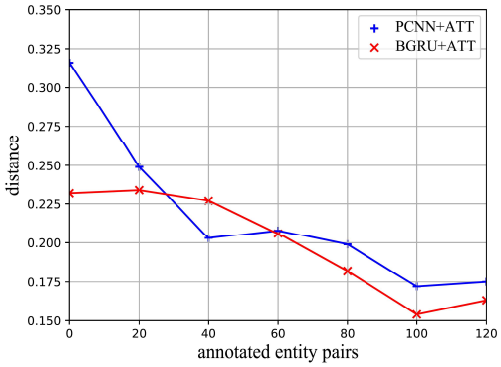

Figure 4: The result of different iterations for the active testing algorithm with PCNN+ATT and BGRU+ATT

our approach, those relation predictions for false negative instances are given a high probability to be corrected. At the same time, true negative instances are accurately identified and given a low (near zero) probability.

C Re-evaluation Discussion

The detailed descriptions and discussions of re-evaluation experiments are conducted in this section.

C.1 Models

PCNN (Zeng et al., 2015) is the first neural method used in distant supervision without human-designed features.

PCNN+ATT (Lin et al., 2016) further integrates a selective attention mechanism to alleviate the influence of wrongly labeled instances. The selective attention mechanism generates attention weights over multiple instances, which is expected to reduce the weights of those noisy instances dynamically.

PCNN+ATT+SL (Liu et al., 2017) is the development of PCNN+ATT. To correct the wrong labels at entity-pair level during training, the labels of entity pairs are dynamically changed according to the confident score of the predictive labels. Clearly, this method highly depends on the quality of label generator, which has great potential to be over-fitting.

PCNN+ATT+RL (Qin et al., 2018b) adopts reinforcement learning to overcome wrong labeling problem for distant supervision. A deep reinforcement learning agent is designed to choose correctly labeled instances based on the performance change of the relation classifier. After that, PCNN+ATT is adopted on the filtered data to do relation classification.

PCNN+ATT+DSGAN (Qin et al., 2018a) is an adversarial training framework to learn a sentence level true-positive generator. The positive samples generated by the generator are labeled as negative to train the generator. The optimal generator is obtained when the discriminator cannot differentiate them. Then the generator is adopted to filter distant supervision training dataset. PCNN+ATT is applied to do relation extraction on the new dataset.

BGRU is one of recurrent neural network, which can effectively extract global sequence information. It is a powerful fundamental model for wide use of natural language processing tasks.

BGRU+ATT is a combination of BGRU and the selective attention.

STPRE (Liu et al., 2018) extracts relation features with BGRU. To reduce inner-sentence noise, authors utilize a Sub-Tree Parse(STP) method to remove irrelevant words. Furthermore, model parameters are initialized with a prior knowledge learned from the entity type prediction task by transfer learning.

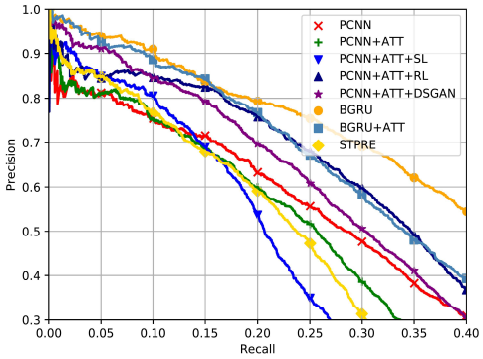

Figure 5: PR curve of distantly supervised relation extractors on NYT-10 with the proposed active testing.

C.2 Discussion

In this section, we additionally provide PR curves to show the performance of baselines. From both Table 3 and Figure 5, we are aware of that: 1) The relative ranking is quite different from that on held-out evaluation according to PR curve. 2) The selective attention has limited help in improving the overall performance, even though it may have positive effects at high confident score. 4) The soft-label method greatly improves the accuracy at high confident score but significantly reduces the overall performance. We deduce that it is severely

	Instances	Real Label	Prediction	Probability
false negative	He renewed that call four years ago in a document jointly written with *Ami_Ayalon*, a former chief of *Israel*'s shin bet security agency and a leader of the labor party.	/person/nationality	/person/nationality	1.0(vetted)
	But, if so, you probably would not be familiar with the town of *Ramapo* in *Rockland_County*.	/location/contain	/location/contain	0.842
	Mr. voulgaris lives in oyster bay but has summered on shelter island since he was a child growing up in *Huntington* in western *Suffolk_County*.	/location/contain	/location/contain	0.837
true negative	His visit opened a new level of debate in *Israel* about the possibility of negotiations with the Syrian president, *Bashar_Al-Assad*.	NA	/person/nationality	0.0(vetted)
	They are in the united states, the *United_Kingdom* and *Canada*, among other places, but not in the Jewish settlements of the west bank.	NA	/administrative_division/country	0.0
	Mr. spielberg and stacey snider, the former *Universal_Pictures* studio chairman who joined *DreamWorks* last year as chief executive, have sole authority to greenlight films that cost $ 85 million or less.	NA	/person/company	0.088

Figure 6: A case study of active testing approach for distantly supervised relation extraction. The entities are labeled in red. $1.0(vetted)$ and $0.0(vetted)$ mean that the entity pair is vetted in our method.

affected by the unbalanced instance numbers of different relations, which will make label generator over-fitting to frequent labels. 4) For the overall performance indicated by PR curves, BGRU is the most solid relation extractor.