A Appendices

A.1 Logistic Regression

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

$$
\begin{aligned}
p(z_i'|y_i', s_i') &= \frac{p(z_i', y_i', s_i')}{\sum_{v \in \{0,1\}} p(z_i' = v, y_i', s_i')} \\
&= \frac{p(z_{jk}, y_{jk}, s_{jk})}{\sum_{v \in \{0,1\}} p(z_{jk} = v, y_{jk}, s_{jk})} \\
&= \frac{p(y_{jk}|z_{jk}, s_{jk}) p(z_{jk}|s_{jk})}{\sum_{v \in \{0,1\}} p(y_{jk}|z_{jk} = v, s_{jk}) p(z_{jk} = v|s_{jk})}
\end{aligned}
$$

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:

$$
p(z_i'|y_i', s_k') = \frac{p(y_{jk}|z_{jk}) p(z_{jk}|s_{jk})}{\sum_{v \in \{0,1\}} p(y_{jk}|z_{jk} = v) p(z_{jk} = v|s_{jk})}
$$

A.2 Vetting Strategy

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

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

A.3 Experimental result of BGRU+ATT

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.

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.

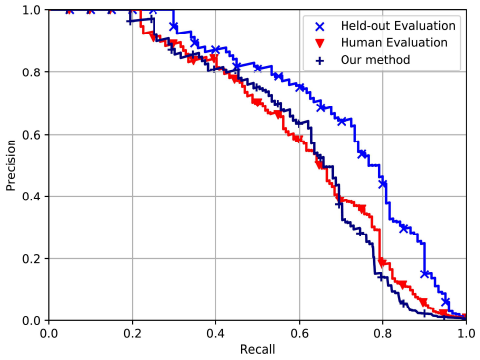

Figure 3: The PR curve of BGRU+ATT on NYT-19.

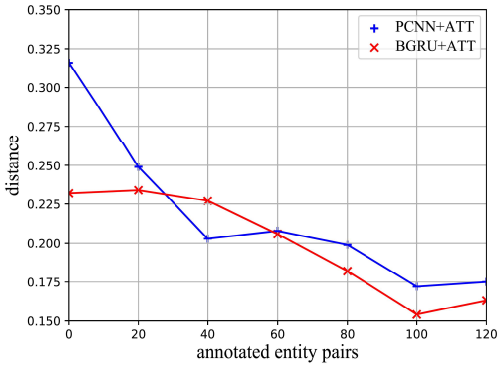

Figure 4: The result of different iterations for the active testing algorithm with PCNN+ATT and BGRU+ATT

B Case Study

We present realistic cases in NYT-10 to show the effectiveness of our method. In Figure 5, 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 our approach, those relation predictions for false negative instances are given a high probability to be corrected. At the same time, true negative in-

	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 5: 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.

stances 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.

C.2 Discussion

In this section, we additionally provide PR curves to show the performance of baselines. From both

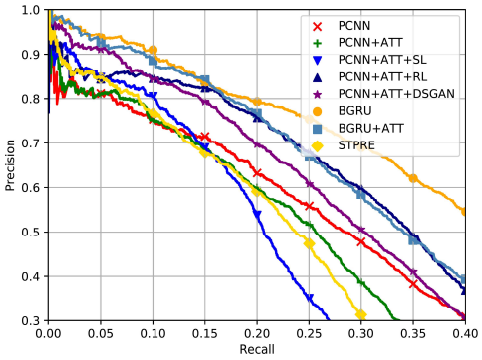

Figure 6: PR curve of distantly supervised relation extractors on NYT-10 with the proposed active testing.

Table 3 and Figure 6, 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 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.