A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction

Tianyu Liu Kexiang Wang Baobao Chang Zhifang Sui

MOE Key Laboratory of Computational Linguistics, Peking University
School of Electronics Engineering and Computer Science, Peking University

{tianyu0421, wkx, chbb, szf}@pku.edu.cn

Abstract

●**Task:**

Distantly supervised relation extraction

●**Problem:**

Distant-supervised relation extraction suffers from **wrong labeling** problems because it heuristically labels relational facts with knowledge bases. Previous denoise model use hard labels which are only determined by distant supervision and immutable during training.

●**Proposal:**

An entity-pair level denoise method which exploits semantic information from correctly labeled entity pairs to correct wrong labels dynamically during training.

Example of Soft-label Corrections during Training

1. False Positive: Place lived → **Place of death**
[***Fernand nault***] , one of canada 's foremost dance figures , died in [***montreal***] on tuesday .

2. False Positive: Place lived → **NA**
[***Alexandra Pelosi***] , a daughter of representative nancy pelosi ..., and paul pelosi of [***san francisco***], was married yesterday to michiel vos.

3. False Negative: NA → **Nationality**
By spring the renowned chef [***Gordon Ramsay***] of [***England***] should be in hotels here.

4. False Negative: NA → **Work In**
..., said [***Billy Ccox***] , a spokesman for [***the United States Department of Agriculture***].

Proposed Model Background

● **Key point**

We propose a joint score function to obtain soft labels during training by taking both the confidence of DS labels and the entity-pair representations into consideration.

● **Model Input**

Entity pair: $\langle h_i, t_i \rangle$

Related sentence: $\{x_1, x_2, \ldots, x_c\}$

Distant-supervised label: L_i

● **Soft-label calculation**

Entity-pair representation: $s_i = \sum_{i=1}^{c} \alpha_i \cdot x_i$

Relation score: $o_t = \frac{\exp(Ms_t + b)}{\sum_k \exp(Ms_k + b)}$

Soft label: $r_i = argmax(o + \max(o) \cdot A \odot L_i)$

● **Training and testing**

Training: $J(\theta) = \sum_{i=1}^{n} \log P(r_i | s_i; \theta)$

Testing: $J(\theta) = \sum_{i=1}^{n} \log P(l_i | s_i; \theta)$

Our method derives a soft label as the gold label for each entity pair dynamically during training, which is not necessarily the same label as the distant supervised (DS) label. We still use DS labels while testing.

Experiments

●Dataset:

New York Times (NYT) corpus

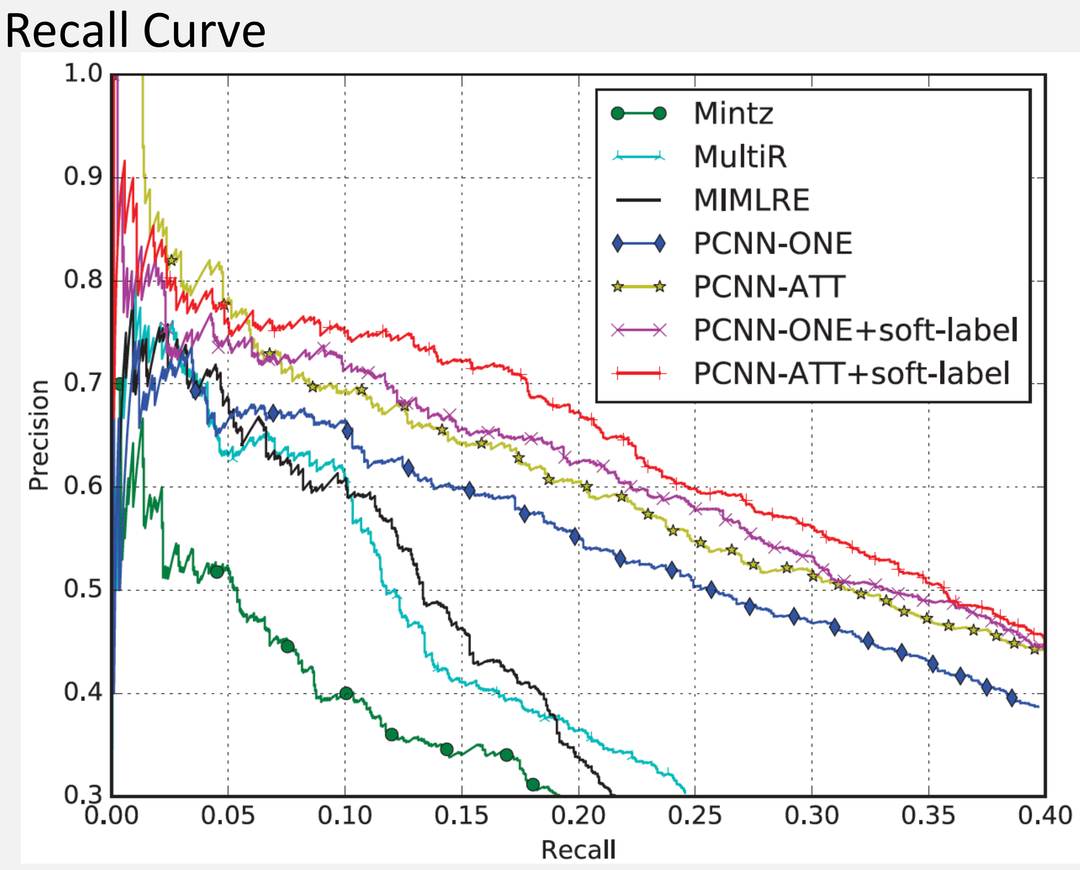
The dataset uses Freebase as distant-supervised knowledge base and New York Times (NYT) corpus as text resource. There are **53** possible relations (including NA).

	# of sentence	# of entity pair	# of facts(no NA)
Train	522611	281270	18252
Test	172448	96678	1950

● Top N precision

Test Settings	One				Two				All			
P@N(%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
PCNN-ONE	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
+soft-label	77.0	72.5	67.7	72.4	80.0	74.5	69.7	74.7	84.0	81.0	74.0	79.7
PCNN-ATT	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
+soft-label	**84.0**	**75.5**	**68.3**	**75.9**	**86.0**	**77.0**	**73.3**	**78.8**	**87.0**	**84.5**	**77.0**	**82.8**

● Precision Recall Curve



MultiR (Hoffmann et al., 2011) and MIMLRE (Surdeanu et al., 2012) are feature-based models. ONE (Zeng et al., 2015) and ATT (Lin et al., 2016) are neural network models based on at-least-one assumption and selective attention, respectively.

Error Analysis: Two Typical Mislabeling

Case 1: Place of Birth → Nationality

[***Marcus Samuelsson***] began … when he was visiting his native [***Ethiopia***].

[***Marcus Samuelsson***] chef born in [***Ethiopia***] and raised in Sweden .

Case 2: Location Contains→ NA

..., he is from neighboring towns in [***Georgia***] (such as Blairsville and [***Young Harris***])

Analysis: Case 1 fail to distinguish similar relations between entities because of their similar sentence patterns. In Case 2, factual relation 'location contains' is mistaken as NA.

Conclusion

This paper proposes a noise-tolerant method to combat wrong labels in distant-supervised relation extraction with soft labels. Our model focuses on entity-pair level noise while previous models only dealt with sentence level noise. Our model achieves significant improvement over baselines on the benchmark dataset. Case study shows that soft-label corrections are of high accuracy.