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

Tianyu Liu, Kexiang Wang, Baobao Chang and **Zhifang Sui**
Key Laboratory of Computational Linguistics, Ministry of Education,
School of Electronics Engineering and Computer Science, Peking University, Beijing, China
{tianyu0421, wkx, chbb, szf}@pku.edu.cn

Abstract

Distant-supervised relation extraction inevitably suffers from wrong labeling problems because it heuristically labels relational facts with knowledge bases. Previous sentence level denoise models don't achieve satisfying performances because they use hard labels which are determined by distant supervision and immutable during training. To this end, we introduce an entity-pair level denoise method which exploits semantic information from correctly labeled entity pairs to correct wrong labels dynamically during training. We propose a joint score function which combines the relational scores based on the entity-pair representation and the confidence of the hard label to obtain a new label, namely a soft label, for certain entity pair. During training, soft labels instead of hard labels serve as gold labels. Experiments on the benchmark dataset show that our method dramatically reduces noisy instances and outperforms the state-of-the-art systems.

1 Introduction

Relation Extraction (RE) aims to obtain relational facts from plain text. Traditional supervised RE systems suffer from lack of manually labeled data. Mintz et al. (2009) proposes distant supervision, which exploits relational facts in knowledge bases (KBs). Distant supervision automatically generates training examples by aligning entity mentions in plain text with those in KB and labeling entity pairs with their relations in KB. If there's no relation link between certain entity pair in KB, it will be labeled as negative instance (NA). However, the automatic labeling inevitably accompanies with wrong labels because the relations of entity pairs might be missing from KBs or mislabeled.

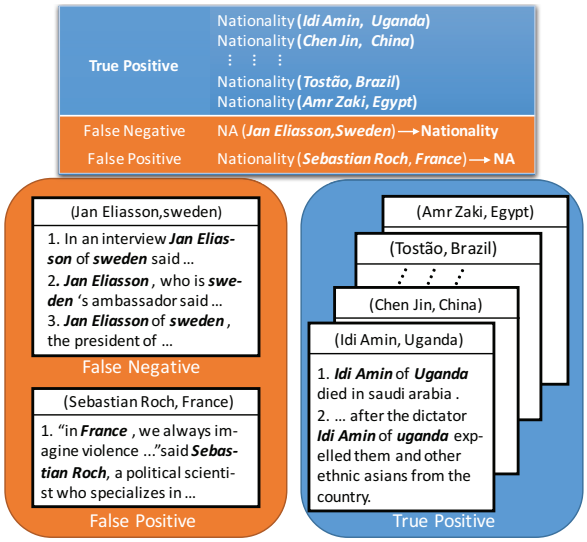

Figure 1: An example of soft-label correction on *Nationality* relation. We intend to use syntactic/ semantic information of correctly labeled entity pairs (blue) to correct the false positive and false negative instances (orange) during training.

Multi-instances learning (MIL) is proposed by Riedel et al. (2010) to combat the noise. The method divides the training set into multiple bags of entity pairs (shown in Fig 1) and labels the bags with the relations of entity pairs in the KB. Each bag consists of sentences mentioning both head and tail entities. Much effort has been made in reducing the influence of noisy sentences within the bag, including methods based on at-least-one assumption (Hoffmann et al., 2011; Ritter et al., 2013; Zeng et al., 2015) and attention mechanisms over instances (Lin et al., 2016; Ji et al., 2017).

However, the sentence level denoise methods can't fully address the wrong labeling problem largely because they use a hard-label method in which the labels of entity pairs are immutable dur-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1790–1795
Copenhagen, Denmark, September 7–11, 2017.

ing training, no matter whether they are correct or not. As shown in Fig 1, due to the absence of (*Jan Eliasson*[1] , *Sweden*) from *Nationality* relation in the KB, the entity pair is mislabeled as NA. However, we find the sentences in the bag of (*Jan Eliasson, Sweden*) share similar semantic pattern *"X of Y"* with correctly labeled instances (blue). In the false positive instance, *Sebastian Roch* is indeed from *France*, but the syntactic pattern of the sentence in the bag differs greatly from those of correctly labeled instances. Actually, the reliability of a distant-supervised (DS) label can be determined by the syntactic/semantic similarity between certain instance and the potential correctly labeled instances. Soft-label method intends to utilize corresponding similarities to correct wrong DS labels in the training stage dynamically, which means the same bag may have different soft labels in different epochs of training. The basis of soft-label method is the dominance of correctly labeled instances. Fortunately, Xu et al. (2013) proves that correctly labeled instances account for 94.4% (including true negatives) in the distant-supervised New York Times corpus (benchmark dataset).

To this end, we introduce a soft-label method to correct wrong labels at entity-pair level during training by exploiting semantic/syntactic information from correctly labeled instances. In our model, the representation of certain entity pair is a weighted combination of related sentences which are encoded by piecewise convolutional neural network (PCNN) (Zeng et al., 2015). Besides, 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. Our contributions are three-fold:

- To the best of our knowledge, we first propose an entity-pair level noise-tolerant method while previous works only focused on sentence level noise.
- We propose a simple but effective method called soft-label method to dynamically correct wrong labels during training. Case study shows our corrections are of high accuracy.
- We evaluate our model on the benchmark dataset and achieve substantial improvement compared with the state-of-the-art systems.

[1]Jan Eliasson is a Swedish diplomat.

2 Methodology

Multi-instances learning (MIL) framework splits the training set $\mathbf{M}$ into multiple entity-pair bags $\{\langle h_1, t_1\rangle, \langle h_2, t_2\rangle, \cdots, \langle h_n, t_n\rangle\}$. Each bag $\langle h_i, t_i\rangle$ contains sentences $\{x_1, x_2, \cdots, x_c\}$ which mention both head entity h_i and tail entity t_i. The representation $\mathbf{s}_i$ of bag $\langle h_i, t_i\rangle$ is a weighted combination of related sentence vectors $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_c\}$ which are encoded by CNN. Finally, we use soft-label score function to correct wrong labels of bags of entity pairs while computing probabilities for each relation type.

2.1 Sentence Encoder

We get the representation of certain sentence $\mathbf{x}_i = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m\}$ by concatenating word embeddings $\{w_1, w_2, \cdots, w_m\}$ and position embeddings (Zeng et al., 2014) $\{p_1, p_2, \cdots, p_m\}$, where $\mathbf{w}_i \in \mathbb{R}^d, w_i \in \mathbb{R}^{d_w}, p_i \in \mathbb{R}^{d_p} (d = d_w + d_p)$.

Convolution layer utilizes a sliding window of size l. We define $\mathbf{q}_i \in \mathbb{R}^{l \times d}$ as the concatenation of words within the i-th window.

$$\mathbf{q}_i = \mathbf{w}_{i-l+1:i} (1 \leq i \leq m + l - 1) \quad (1)$$

The convolution matrix is denoted by $\mathbf{W}_c \in \mathbb{R}^{d_c \times (l \times d)}$, where d_c is the sentence embedding size. The i-th filter of the convolutional layer is computed as:

$$\mathbf{f}_i = [\mathbf{W}_c \mathbf{q} + \mathbf{b}]_i \quad (2)$$

Afterwards, Piecewise max-pooling (Zeng et al., 2015) is used to divide convolutional filter $\mathbf{f}_i$ into three parts $\{\mathbf{f}_i^1, \mathbf{f}_i^2, \mathbf{f}_i^3\}$ by head and tail entities. For example, the sentence "*Barack Obama* was born in *Honululu* in 1961" are divided into '*Barack Obama*', 'was born in *Honululu*' and 'in 1961'. We perform max-pooling on these three parts separately, and the i-th element of sentence vector $\mathbf{x} \in \mathbb{R}^{d_c}$ is defined as the concatenation of them:

$$\mathbf{x}_i = [max(\mathbf{f}_i^1); max(\mathbf{f}_i^2); max(\mathbf{f}_i^3)] \quad (3)$$

2.2 Sentence Level Weight distribution

The representation of entity pair $\langle h_i, t_i\rangle$ is defined as a weighted combination of sentences in the bag.
At-least-one: At-least-one assumption is a down sampling method which assumes at least one sentence in the bag will express the relation between two entities, and select the most likely sentence in the bag for training and prediction. To be more

specific, the weight of the selected sentence is 1 while those of other sentences in the bag are all 0.

Selective Attention: Lin et al. (2016) proposes selective attention mechanism to reduce weights of noisy instances within the entity-pair bag.

$$\mathbf{s} = \sum_i \alpha_i \mathbf{x}_i; \alpha_i = \frac{\exp(\mathbf{x}_i \mathbf{A} \mathbf{r})}{\sum_k \exp(\mathbf{x}_k \mathbf{A} \mathbf{r})} \quad (4)$$

where α_i is the weight of sentence vector $\mathbf{x}_i$, $\mathbf{A}$ and $\mathbf{r}$ are diagonal and relation query parameters.

2.3 Soft-label Adjustment

The key of our method is to derive a soft label as the gold label for each bag dynamically during training, which is not necessarily the same label as the distant supervised (DS) label. We still use DS labels while testing.

The soft label is determined dynamically, which means the same bag may have different soft labels in different training epochs. we propose following joint function to determine the soft label r_i for entity pair $\langle h_i, t_i \rangle$:

$$r_i = \arg\max(\mathbf{o} + \max(\mathbf{o})\mathbf{A} \odot L_i) \quad (5)$$

where $\mathbf{o}, \mathbf{A}, L_i \in \mathbb{R}^{d_r}$ (d_r is the number of predefined relations). One-hot vector L_i indicates the DS label of $\langle h_i, t_i \rangle$. Relation Confidence vector $\mathbf{A}$ represents the reliability of DS labels. Each element in $\mathbf{A}$ is a decimal between 0 and 1, which indicates the confidence of corresponding DS labeled relation type. $\odot$ operation represents element-wise production. $\mathbf{o}$ is the vector of relational scores based on the entity-pair representation $\mathbf{s}_i$ of $\langle h_i, t_i \rangle$. $\max(\mathbf{o})$ is a scaling constant which restricts the effect of the DS label. The score of the t-th relation type $\mathbf{o}_t$ is calculated based on the trained relation matrice $\mathbf{M}$ and bias $\mathbf{b}$:

$$\mathbf{o}_t = \frac{\exp(\mathbf{M}\mathbf{s}_t + \mathbf{b})}{\sum_k \exp(\mathbf{M}\mathbf{s}_k + \mathbf{b})} \quad (6)$$

We use entity-pair level cross-entropy loss function using soft labels as gold labels while training:

$$J(\theta) = \sum_{i=1}^{n} \log p(r_i | \mathbf{s}_i; \theta) \quad (7)$$

In the testing stage, we still use the DS label l_i of certain entity pair $\langle h_i, t_i \rangle$ as the gold label:

$$G(\theta) = \sum_{i=1}^{n} \log p(l_i | \mathbf{s}_i; \theta) \quad (8)$$

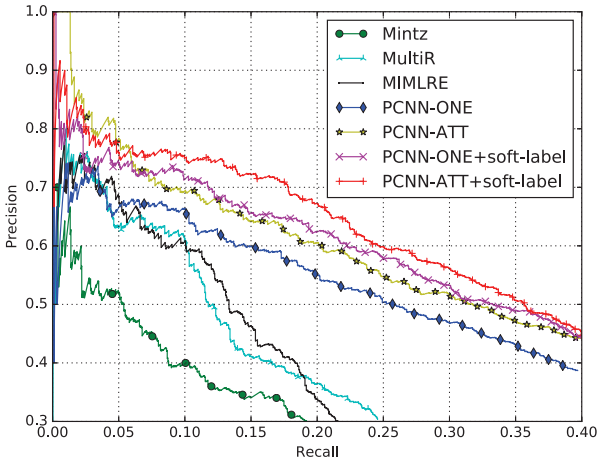

Figure 2: Precision/Recall curves of our model and previous state-of-the-art systems. Mintz (Mintz et al., 2009), MultiR (Hoffmann et al., 2011) and MIMLRE (Surdeanu et al., 2012) are feature-based models. ONE (Zeng et al., 2015) and ATT (Lin et al., 2016) are neural network models based on at-least-one assumption and selective attention, respectively.

3 Experiments

In this section, we first introduce the dataset and evaluation metrics in our experiments. Then, we demonstrate the parameter settings in our experiments. Besides, we compare the performance of our method with state-of-the-art feature-based and neural network baselines. Case study shows our soft-label corrections are of high accuracy.

3.1 Dataset and Evaluation Metrics

We evaluate our model on the benchmark dataset proposed by Mintz et al. (2009), which has also been used by Riedel et al. (2010), Hoffmann et al. (2011), Zeng et al. (2015) and Lin et al. (2016). The dataset uses Freebase (Bollacker et al., 2008) as distant-supervised knowledge base and New York Times (NYT) corpus as text resource. Sentences in NYT of the years 2005-2006 are used as training set while sentences in NYT of 2007 are used as testing set. There are 53 possible relations including NA which indicates no relation. The training data contains 522611 sentences, 281270 entity pairs and 18252 relational facts. The testing data contains 172448 sentences, 96678 entity pairs and 1950 relational facts.

Similar to the previous work, We report both aggregate precision/recall curves and top-N precision (P@N).

Window size	Word dimension	Position dimension	Filter dimension	Batch size	Learning rate	Dropout
$l = 3$	$d_w = 50$	$d_p = 5$	$d_c = 230$	$B = 160$	$\lambda = 0.001$	$p = 0.5$

Table 1: Parameter settings of our experiments.

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**

Table 2: Top-N precision (P@N) for relation extraction in the entity pairs with different number of sentences. Following (Lin et al., 2016), One, Two and All test settings random select one/two/all sentences on the bags of entity pairs from the testing set which have more than one sentence to predict relation.

3.2 Comparison with previous work

Mintz (Mintz et al., 2009), MultiR (Hoffmann et al., 2011) and MIMLRE (Surdeanu et al., 2012) are feature-based models. PCNN-ONE (Zeng et al., 2015) and PCNN-ATT (Lin et al., 2016) are piecewise convolutional neural network (PCNN) models based on at-least-one assumption and selective attention, which are introduced in Section 2.2, respectively. All the results of compared models come from the data reported in their papers.

3.3 Experimental Settings

We use cross-validation to determine the parameters in our model. Soft-label method uses PCNN-ONE/PCNN-ATT to represent the bags of entity pairs, and we don't tune on the parameters of PCNN-ONE/PCNN-ATT for fair comparsion. So we use the same pre-trained word embeddings and parameters of CNN encoder as those of Lin et al. (2016). Detailed parameter settings are shown in Table 1. Moreover, we use Adam optimizer. Besides, to avoid negative effects of dominant NA instances in the begining of training, soft-label method is adopted after 3000 steps of parameter updates. The confidence vector $\mathbf{A}$ is heuristically set as $[0.9, 0.7, \cdots, 0.7]$ (the confidence of NA is 0.9 while confidence of other relations are all 0.7).

3.4 Precision Recall Curve

We have following observations from Figure 2: (1) For both ATT and ONE configuration, soft-label method achieves higher precision than baselines when recall is greater than 0.05. After manual evaluation, we find that most wrong instances with less than 0.05 recall are wrong labeling entity pairs in test set. (2) Even weaker baseline PCNN-ONE using soft labels gets a slightly better performance than PCNN-ATT. (3) When recall is between 0.05 and 0.15, the curve of our model ATT+soft-label is relatively stable, which demonstrates soft-label can obtain relatively stable performance in extracting relational facts.

False positive: Place lived → **Place of death**
Fernand nault , one of canada 's foremost dance figures , died in ***montreal*** on tuesday .
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.
False Negative: NA → **Nationality**
By spring the renowned chef ***Gordon Ramsay*** of ***England*** should be in hotels here.
False Negative: NA → **Work in**
⋯, said ***Billy Ccox*** , a spokesman for the ***United States Department of Agriculture***.

Table 3: Some examples of soft-label corrections while training

3.5 Top N precision

Table 2 shows top-N precision (P@N) of the state-of-the-art systems and our model. We can see that (1) For both PCNN-ONE and PCNN-ATT model, soft-label method improves the precisions by over 10% in all test settings, which demonstrates the effects of our model. (2) Even a weaker baseline (PCNN-ONE) with soft-label method achieves higher precision than a strong model (PCNN-ATT). It shows that entity-pair level denoise model performs much better than the models which only focus on sentence level noise.

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***)

Table 4: Two typical wrong corrections of soft-label adjustment during training.

3.6 Case Study

Some examples of soft-label corrections during training are shown in Table 3. We can see that soft-label method can recognize both false positives and false negatives during training and correct wrong labels successfully. The two sentences above are mislabeled as *place lived* because triple facts *(Fernand nault, place lived, Montreal)* and *(Alexandra pelosi, place lived, San francisco)* exist in Freebase. However, the two sentences fail to express *place lived* relation. Our model can automatically correct them by soft-label adjustment. The two sentences below show that our model can also change false negative (NA) examples caused by missing facts in Freebase to correct ones.

Besides, our model has strong ability to distinguish different relational patterns, even for similar relations like *Place lived, Place of born, Place of Death*.

4 Error Analysis

We randomly select 200 instances of soft-label corrections during training for PCNN-ONE and PCNN-ATT respectively and check them manually. The accuracy of soft-label corrections for PCNN-ONE is 88.5% (177/200) while that for PCNN-ATT is 92% (184/200). We notice that the accuarcy of PCNN-ATT+soft-label is slightly higher than that of PCNN-ONE+soft-label. The condition is the same as our expectation. As explained in Sec 2.2, PCNN-ATT has better bag representations than PCNN-ONE because it can reduce the effect of noisy instances within the bag. The soft-label of certain bag is determined by its bag representation and the confidence of corresponding DS label. So the accuracy of soft-label corrections for PCNN-ATT can benefit from better bag representations.

Although most of soft-label corrections are of high accuracy (90.25%), there are still several wrong corrections. Table 4 lists two typical wrong corrections during training. Wrong corrections like Case 1 fail to distinguish similar relations (both *Nationality* and *place of birth* are relations between people and locations) between entities because of their similar sentence patterns. However, wrong corrections like Case 1 are rare (5/39) in our experiments. Soft-label method can still distinguish most similiar relations as shown in Sec 3.6. In Case 2, factual relation *location contains* is mistaken as NA partially because the relational pattern of this sentence is somewhat different from the regular *location contains* pattern. Additionally, soft-label method has a tendency to label ambiguous facts as NA because negative instances (NA) are dominated in the corpus. However, most bags which are soft-labeled as NA are still well-labeled in our experiments.

We argue that the minor wrong corrections of relational facts during training don't affect the overall performance much because distant supervision doesn't lack instances of relational facts due to its strong ability to automatically label large web text.

5 Conclusion and Future Work

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.

In the future, we plan to develop a new measurement for the reliability of certain distantly supervised label by evaluating the corresponding similarity between certain instance and the potential correctly labeled instances instead of using heuristically set confidence vector. In addition, we tend to find a more suitable sentence encoder rather than piece-wise CNN for our soft-label method.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work is supported by the National Key Basic Research Program of China (No. 2014CB340504) and the National Natural Science Foundation of China (No.61375074, 61273318). The contact authors are Zhifang Sui and Baobao Chang.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3060–3066.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, volume 1, pages 2124–2133.

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 ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378.

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*, pages 455–465. Association for Computational Linguistics.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *ACL (2)*, pages 665–670.

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), Lisbon, Portugal*, pages 17–21.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.