Cooperative Denoising for Distantly Supervised Relation Extraction

Kai Lei[1,*], **Daoyuan Chen**[1,*], **Yaliang Li**[2], **Nan Du**[2], **Min Yang**[3], **Wei Fan**[2], **Ying Shen**[1,†]
[1]School of Electronics and Computer Engineering, Peking University Shenzhen Graduate School
[2]Tencent Medical AI Lab
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[1]{leik, shenying}@pkusz.edu.cn, [1]chendaoyuan@pku.edu.cn
[2]{yaliangli,ndu,davidwfan}@tencent.com, [3]min.yang@siat.ac.cn

Abstract

Distantly supervised relation extraction greatly reduces human efforts in extracting relational facts from unstructured texts. However, it suffers from noisy labeling problem, which can degrade its performance. Meanwhile, the useful information expressed in knowledge graph is still underutilized in the state-of-the-art methods for distantly supervised relation extraction. In the light of these challenges, we propose CORD, a novel **CO**ope**R**ative **D**enoising framework, which consists two base networks leveraging text corpus and knowledge graph respectively, and a cooperative module involving their mutual learning by the adaptive bi-directional knowledge distillation and dynamic ensemble with noisy-varying instances. Experimental results on a real-world dataset demonstrate that the proposed method reduces the noisy labels and achieves substantial improvement over the state-of-the-art methods.

1 Introduction

Relation extraction aims to discover the semantic relationships between entities. Recently, it has attracted increasing attention due to its broad applications in many machine learning and natural language processing tasks such as Knowledge Graph (KG) construction (Shin et al., 2015), information retrieval (Kadry and Dietz, 2017), and question answering (Abujabal et al., 2017).

Distant supervision is one of the most important techniques in practice for relation extraction due to its ability to generate large-scale labeled training data automatically by aligning KGs to text corpus. Despite its effectiveness, it suffers from noisy labeling problem such as false negative examples, which can severely degrade its performance. To alleviate this limitation, multi-instance learning and probabilistic graphical models have been widely explored by existing work (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012). Due to the success of deep learning, there has been increasing interest in applying deep neural networks to solve this problem. Zeng et al. (2015) proposed a piecewise CNN model combining multi-instance paradigm, Lin et al. (2016) and Lin et al. (2017) introduced sentence-level and multi-lingual attention to alleviate the side-effect caused by noisy instances, and Luo et al. (2017) further modeled noises explicitly.

However, most existing studies reduce noises by leveraging information within text corpus, ignoring the relational facts expressed in other information sources, such as KG triples or semi-structured tables. Leveraging various information sources simultaneously, which takes full advantage of diverse and supplementary information in different sources, is beneficial to reduce noisy labels for distant supervision. To be more specific, we transform each sentence into entity sequence based on KG information, which is helpful to locate critical entities and adjust word-based network when their predictions are not consistent.

Moreover, by incorporating information from other sources, distantly supervised relation extraction methods can better handle "Not A relation" (NA) class, which is the main reason for the noisy label problem. The large proportion of NA instances is typical in distantly supervised relation extraction task, and it's non-trivial to characterize the NA patterns only based on text-corpus information. By considering

* Equal contribution.
† Corresponding author.

Proceedings of the 27th International Conference on Computational Linguistics, pages 426–436
Santa Fe, New Mexico, USA, August 20-26, 2018.

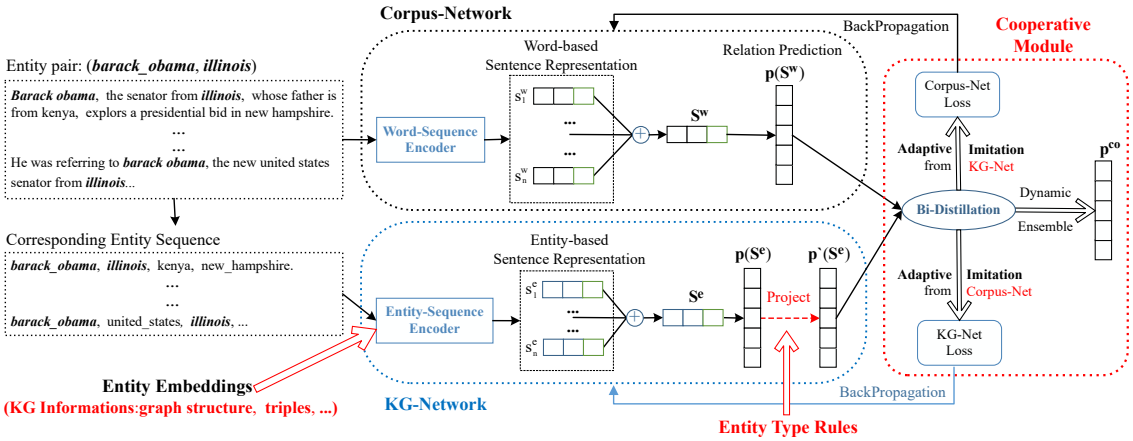

Figure 1: Overview of the proposed Cooperative Denoising Framework.

the information from KG, entity sequence can easily discriminate NA from other relations, since entities appeared in NA sentence usually are not connected in KG.

Motivated by the above observations, we propose a novel cooperative denoising framework (see Figure 1) to leverage the corpus-based and KG-based information. Specifically, we design two base networks, Corpus-Net and KG-Net, which are modeled with two separate Gated Recurrent Unit (GRU) networks, to predict relations using word-sequence and entity-sequence respectively. For KG-Net, we employ network embedding and KG embedding methods to pre-train the entity embeddings, and then project the prediction to a logic rule regularization subspace. Afterward, we design a cooperative module which involves the interactive learning between the two base networks with an adaptive bi-directional knowledge distillation mechanism, and the predictions of the base networks are dynamically integrated by an ensemble method. The key insight is that the base networks trained on different sources can learn complementary information, and thus the cooperative learning can benefit from the complementarity of different expressions of the same relational fact.

Our main contributions are as follows:

- We explore the feasibility of distantly supervised relation extraction by leveraging the information from different sources cooperatively.
- We devise a bi-directional knowledge distillation mechanism to enhance each base network via supplementary supervision.
- We design an adaptive imitation rate setting and a dynamic ensemble strategy to guide the training procedure and help the prediction of noisy-varying instances.
- The experimental results on a benchmark dataset show that the proposed method has robust superiority over compared methods.

2 Methodology

In this paper, we focus on the task of distantly supervised relation extraction. Our goal is to predict relation r for a given entity pair $< e_1, e_2 >$. The proposed framework CORD conducts with multi-instance learning, i.e., we take a bag of sentences mentioning both entity e_1 and e_2 as input, and we compute the probabilities for each relation expressed by this bag as output.

As Figure 1 shows, given a collection of sentences containing the target entity pair, we first transform each sentence into its distributed word-sequence and entity-sequence representations, and predict relation respectively using the attention weighted representation via a multi-instance learning mechanism. We also project the prediction of KG-Net to a logic rule regularization subspace. Then, we train the two base networks simultaneously with a bi-directional knowledge distillation method, in which the predictions of KG-Net and Corpus-Net are used as soft labels for each other. The final prediction is the ensemble

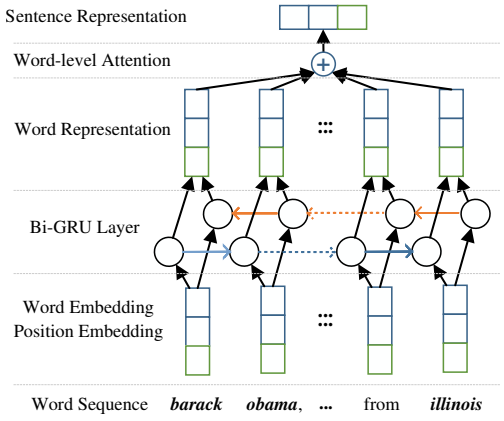

Figure 2: Word-Sequence Encoder.

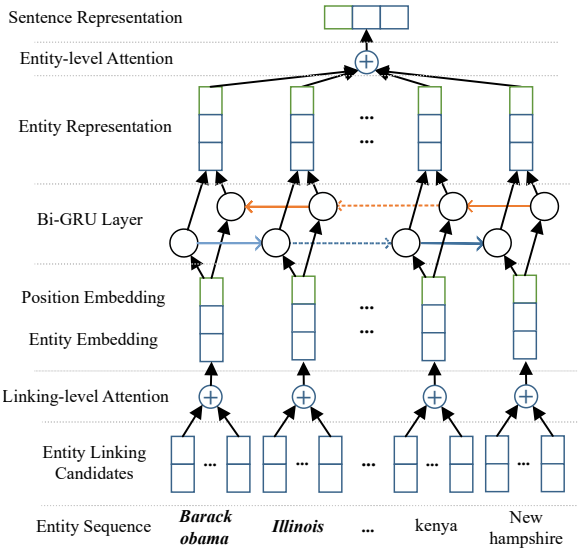

Figure 3: Entity-Sequence Encoder.

of the two base networks, and their weights in the ensemble are dynamically adjusted. In the rest of this section, we elaborate each component of CORD in detail.

2.1 Corpus-Network

We first introduce the word-sequence encoder shown in Figure 2, which transforms each sentence s_i to corresponding word-based sentence representation $\mathbf{s_i^w}$.

Input Representation Each word in the sentence is mapped to a low-dimensional embedding $w_i \in \mathbb{R}^{d_w}$ through a word embedding layer, where d_w denotes the size of the embedding. Similar to Zeng et al. (2014) and Lin et al. (2016), we encode the relative distances to entity e_1 and e_2 as $p_i \in \mathbb{R}^{2d_p}$, where $2d_p$ is the size of the position embedding. The distances are helpful in emphasizing how informative the word is thereby enabling better discrimination for relation extraction. We concatenate word vector w_i and position vector p_i as $\mathbf{w_i} = w_i \| p_i$, $\mathbf{w_i} \in \mathbb{R}^{d_w+2d_p}$, and feed words input $\mathbf{s_i} = \{\mathbf{w_1}, \ldots, \mathbf{w_n}\}$ to Bi-GRU layer.

Bi-directional GRU Layer We employ bi-directional GRU to encode patterns in $\mathbf{s}_i$ into a hidden representation $\mathbf{s_i^w} = \{h_1^f \| h_T^b, \ldots, h_T^f \| h_1^b\}$, which is obtained by concatenating each forward state h_j^f and backward state h_{T-j+1}^b at step j given input $\mathbf{s}_i$, $\mathbf{s_i^w} \in \mathbb{R}^{2d_h}$ and d_h means the hidden size of GRU. The forward sequence $[h_1^f, h_2^f, \ldots, h_T^f]$ and backward sequence $[h_1^b, h_2^b, \ldots, h_T^b]$ are calculated by GRU (Cho et al., 2014).

Hierarchical Attention As Figure 2 and Figure 1 show, we apply hierarchical attention for Corpus-Net involving word and sentence levels respectively. The motivation is that the semantic meaning of a sentence bag representation is gathered by vectors in different levels from the bottom to up, and the vectors in different context do not contribute equally. Some words express targeted relation more relevantly than others in a sentence. Furthermore, since we concatenate position vector to each word, a word with different distance is also importance-varying. As for sentence-level attention, it can reduce the weights of the sentences that suffer from wrong-labeling and improper-bagging (i.e., express inconsistent relations) problems. Inspired by Lin et al. (2016), we calculate the aggregation of different level attentions with unified form as:

$$\mathbf{X} = \sum_{i=1}^{n} a_i \mathbf{x}_i; \qquad a_i = \frac{exp(\mathbf{x_i A r})}{\sum_{j=1}^{n} exp(\mathbf{x_j A r})}, \tag{1}$$

where $\mathbf{x_i}$ is individual input vector in different levels, $\mathbf{X}$ is the corresponding sum of them, $\mathbf{r}$ is randomly initialized global relation vector. Note that we set different $\mathbf{r}$ and weighted diagonal matrix $\mathbf{A}$ for different attention levels.

Finally, the Bi-GRU output of each word $\mathbf{w_i}$ is gathered to sentence representation $\mathbf{s^w}$, and then to the bag-wise representation $\mathbf{S^w}$. We feed the resulting vector $\mathbf{S^w}$ to a *Softmax* classifier.

Prediction With $\mathbf{S^w}$ as input, the condition probability of each relation j is calculated as:

$$\mathbf{p_j}(\mathbf{S^w}) = \frac{\exp(o_j)}{\sum_{i=1}^{n_r} \exp(o_i)}; \qquad \mathbf{o} = \mathbf{W}\mathbf{S^w} + \mathbf{b}, \tag{2}$$

where $\mathbf{o} = [o_1, \ldots, o_{n_r}]$ is calculated with coefficient matrix $\mathbf{W} \in \mathbb{R}^{n_r \times (2d_h)}$ and bias $\mathbf{b} \in \mathbb{R}^{n_r}$. n_r is the number of relations and o_j measures how well $\mathbf{S^w}$ matches relation j.

2.2 KG-Network

Besides Corpus-Net, we propose another base network to incorporate information of KG. The entity-sequence encoder transforms each sentence s_i to entity-based sentence representation $\mathbf{s_i^e}$ as illustrated in Figure 3.

Input Representation One of the key challenges of leveraging KG information is identifying what information to use. Here we employ an extensible manner by using different entity embedding methods. For instance, *network embedding* methods such as DeepWalk (Perozzi et al., 2014) primarily encode graph structure information, while *knowledge embedding* methods such as TransE (Bordes et al., 2013) usually focus on triples information, we can flexibly use one of them or merge them together. Specifically, we link the detected entity names in sentences to the Freebase5M (Bordes et al., 2015) by n-gram text matching, and use the DeepWalk and TransE embeddings of linked entity candidates denoted as $\{e_{1,1}, e_{1,2}, \ldots, e_{m,k-1}, e_{m,k}\}$, where k is the amount of candidates for each entity, m is the amount of entities appearing in sentence.

In addition, we use the position vectors of Corpus-Net for KG-Net because the word-based distances are more discriminative than the entity-based distances. To be specific, the transformation between them is not one-to-one mapping because different word-sequences may result in the same entity-sequence and hence loss of information, e.g., "Obama flew to the US" and "Obama left the US" are both mapped to "Obama, US".

Bi-GRU and Attention Then we employ a similar architecture of the Bi-GRU component and the attention aggregation of Corpus-Net. As shown in Figure 3, the linked candidates $\{e_{i,1}, \ldots, e_{i,k}\}$ are aggregated to e_i for each entity, then entity vectors $\{e_1, \ldots, e_m\}$ and position vectors $\{p_1, \ldots, p_m\}$ are concatenated in element-wise and as input to Bi-GRU layer. The Bi-GRU outputs are gathered to $\mathbf{s^e}$ with entity-level attention, then to $\mathbf{S^e}$ with bag-level attention, and fed to a *Softmax* classifier similar to Corpus-Net.

External Rule Knowledge We regard the patterns learned automatically from word and entity sequences as internal knowledge and manage to transfer external human knowledge such as logic rules into the base network. Furthermore, the KG-specific rules can be incorporated into Corpus-Net gradually by bi-distillation and vice versa.

Here we concentrate on relation-specific type rules because: (1) We observe some typical false predictions which could be corrected with type rules (e.g., it is unreasonable to predict two person entities as relation *place of birth* whose tail type cannot be a *person*). (2) We can automatically obtain large-scale type resources because most KG reserved this information. For example, in Freebase, we can collect the types of entities located in *type/instance* field and the relation-specific type constraints located in *rdf-schema#domain* and *rdf-schema#range* fields.

To be specific, given all types (an entity can have many types) for an entity pair j as $t_{j,1}$ and $t_{j,2}$, we design the logic rule for each relation i as $T_{i,1}\ and\ T_{i,2}\ are\ not\ missing \implies (T_{i,1} \in t_{j,1}) \wedge (T_{i,2} \in t_{j,2})$ where $T_{i,1}$ and $T_{i,2}$ are the relation-specific type constrains for relation i. Here we apply probabilistic soft logic (Bach et al., 2017) to encode rules flexibly, i.e., the rule scores are continuous truth values in the internal $[0, 1]$ rather than $\{0, 1\}$ and logic $\wedge$ denotes averaging of truth values.

Moreover, considering type granularity, i.e., type information that is usually specified with hierarchy form (e.g., /people/person/spouse), we divide the number of matched levels (fields) of the type hierarchy by field amount as the value of $(T_i \in t_j)$ to gain fine-grained features.

Finally, with prediction $\mathbf{p}(\mathbf{S^e})$ from *Softmax* classifier, we use a posterior regularization fashion (Hu et al., 2016) to project $\mathbf{p}(\mathbf{S^e})$ into a constrained subspace $\mathbf{p}'(\mathbf{S^e})$ as follows:

$$(\mathbf{p}'(\mathbf{S^e}))^* \propto \mathbf{p}(\mathbf{S^e}) \times \mathrm{e}^{s_r}, \qquad s_r = -\sum_{i=1}^{l} C\lambda_i(1 - r_i(\mathbf{S^e})), \tag{3}$$

where λ_i is confidence of each rule, C is regularization parameter and s_r is rule score factor, which indicates how $\mathbf{S^e}$ satisfies the rules. This is the closed-form solution obtained by solving an optimization problem which finds the optimal $\mathbf{p}'(\mathbf{S^e})$ fitting rules meanwhile staying to $\mathbf{p}(\mathbf{S^e})$. We set λ_i as $\mathbf{p}_i(\mathbf{S^e})$, i.e., the higher probability classifier predicts (believes), the stronger effect rule-constraint takes for relation i. We can design other rules and enable to scale with similar manner.

2.3 Cooperative Module

In this section, we introduce how to ensemble two base networks cooperatively.

Bi-directional Knowledge Distillation We observe that KG-Net and Corpus-Net have different hard examples and different wrong predictions, i.e., for the same sentence bag, Corpus-Net may predict higher probability than KG-Net and sometimes, the contrary (we demonstrate the differences in Experiments Section). This observation encouraged us to train them cooperatively with mutual knowledge supplementation.

We devise a bi-directional knowledge distillation method to enhance their supervision information in label space. Specifically, the two base networks learn with the hard label $\mathbf{y}$ from distant supervision. Meanwhile, we set the predicted probability of the two base networks $\mathbf{p}^c$ and $\mathbf{p}^k$ as soft label to each other simultaneously:

$$L_c = \sum_{i=1}^{N}(\ell(\mathbf{y}_i, \mathbf{p}_i^c) + \pi_c \ell(\mathbf{p}_i^k, \mathbf{p}_i^c)), \qquad L_k = \sum_{i=1}^{N}(\ell(\mathbf{y}_i, \mathbf{p}_i^k) + \pi_k \ell(\mathbf{p}_i^c, \mathbf{p}_i^k)), \tag{4}$$

where ℓ is cross entropy loss, π is imitation rate, and N is batch size. We update the model parameters by minimizing L_c and L_k with Adam (Kingma and Ba, 2014) optimizer.

The learning process can be regarded as the fact that the two base networks not only learn from the coarse-grained hard label which is one-hot and low entropy, but also learn from the teacher network which expresses specific supplementary knowledge and dependencies between relations with soft label. For example, label $[0.3, 0.2, 0.9]$ is more informative than $[0, 0, 1]$.

Also note that the early base network is not reliable and gives low quality knowledge through soft label, so we pre-train the two base networks separately with certain steps before mutual learning.

Adaptive Imitation Rate The classification difficulty of the two base networks is varying with different entity pair bag instance, sometimes KG-Net is a better qualified teacher for Corpus-Net and sometimes vice versa. To transfer more reliable knowledge of each base network to another and train them more effectively, we set the imitation weights as following:

$$\pi_c = \frac{\ell(\mathbf{y}_i, \mathbf{p}_i^k)}{\ell(\mathbf{y}_i, \mathbf{p}_i^c) + \ell(\mathbf{y}_i, \mathbf{p}_i^k)}, \qquad \pi_k = \frac{\ell(\mathbf{y}_i, \mathbf{p}_i^c)}{\ell(\mathbf{y}_i, \mathbf{p}_i^c) + \ell(\mathbf{y}_i, \mathbf{p}_i^k)}, \tag{5}$$

where π_c and π_k are inversely proportional to the hard-label loss of each other, i.e., the smaller the loss is, the more qualified is the base network as teacher toward each other. In addition, from the perspective of optimization, the adaptive imitation can prevent the gradient from being dominated by ill-classified examples and hence be able to train the model effectively.

Later, in Section 3.4, we will demonstrate the effectiveness of the adaptive imitation rate by comparing with the fixed setting.

Dynamic Ensemble Prediction The final prediction $\mathbf{p}^{co}$ of the CORD framework is an ensemble of the two base networks predictions $\mathbf{p}^c$ and $\mathbf{p}^k$ because each of them has its strong points. We propose a dynamic ensemble strategy considering that (1) A high type-rule score indicates KG-Net may classify current sentence bag well because the predictions of the classifier satisfy the rules; (2) Ideally, entity

name in a sentence should be linked to only one entity in KG, so the KG-Net would be more confused with more linked candidates. Thus, with the above two factors, we can specify the dynamic ensemble prediction $\mathbf{p}^{co}$ as follows:

$$\mathbf{p}^{co} = (1 - w_k)\mathbf{p}^c + w_k\mathbf{p}^k, \qquad w_k = \alpha + \beta(\frac{s_r}{n_r} - \frac{n_c}{n_e \times N_e}), \tag{6}$$

where α is the empirical KG-Net base weight, β is the wave range. They can be set as the ratio of some evaluation indicators (such as F-score) of separate-trained base network. Then the prediction weight of KG-Net $w_k \in [\alpha - \beta, \alpha + \beta]$ depends on the normalized ($\in [0, 1]$) rule score factor and candidates score, i.e., average rule score per-relation and number of candidates per-entity dividing N_e, which is the upper limit on the number of candidates for linking. And s_r is the rule score factor in Eq. 3, n_r, n_c n_e are the amounts of relations, all entities candidates and gathered entities in sentence respectively.

As a comparison, we also deploy a naive baseline using static ensemble weight and report the results in Section 3.4.

3 Experiments

In this section, we aim to evaluate the effectiveness of the proposed CORD framework. We conduct an overall performance comparison with baseline methods and perform a comprehensive examination of the KG-Net and the Cooperative Module.

3.1 Experiment setup

Dataset and Evaluation Metrics We conduct experiments on the widely used benchmark dataset NYT10 (Riedel et al., 2010), which is built by aligning triples in Freebase to the New York Times corpus and contains 53 relations. There are 522,611/172,448 sentences, 281,270/96,678 entity pairs, and 18,252/1,950 relation mentions in train/test dataset respectively. Following previous works (Mintz et al., 2009; Lin et al., 2016), we evaluate our method in the held-out evaluation with P-R curve and P@N metric without expensive human evaluations.

Parameter Settings We set the embedding dimensions as 5, 50, 64, 64 for position, word2vec, DeepWalk and TransE respectively. For both base networks, we set the cell size of GRU as 230, learning rate as 0.001, dropout probability as 0.5 and batch size as 20. For the cooperative module, we set the base weight α and wave range β as 0.4, 0.2 respectively, and fixed w_k as 0.4 for ensemble comparison.

3.2 Comparison with Baseline Methods

We compare our approach with three traditional feature-based methods and two state-of-art neural-based methods.

Feature-based Methods *Mintz* (Mintz et al., 2009) is a multiclass logistic regression model; *MultiR* (Hoffmann et al., 2011) is a probabilistic graphical model which can handle overlapping relations; *MIML* (Surdeanu et al., 2012) is also a probabilistic graphical model but using a multi-instance multi-label paradigm.

Neural-based Methods *CNN+ATT* (Lin et al., 2016) is a sentence-level attention model based on CNN, which can dynamically reduce the weights of noisy instances; *PCNN+ATT* (Lin et al., 2016) achieves state-of-art results by applying sentence-level attention to the piecewise max pooling model, PCNN (Zeng et al., 2015).

The precision-recall curve results are shown in Figure 4, where *Base-Net-Corpus* and *Base-Net-KG* are our best results for the two base networks with independent training, *CORD* is the cooperative training and dynamic ensemble results. For the aforementioned five methods, we directly use the results reported in (Lin et al., 2016).

Figure 4 demonstrates that: (1) The KG-Net achieves higher coverage than the feature-based methods, comparable precision with the *MultiR* and *MIML* and an obvious gap with other neural-based methods. This indicates that the KG-Net and feature-based methods can capture certain patterns effectively but with relatively low coverage. On one hand, the decent precision shows potentials of the KG-Net to capture patterns with entity-sequence and KG information. On the other hand, we suggest that the

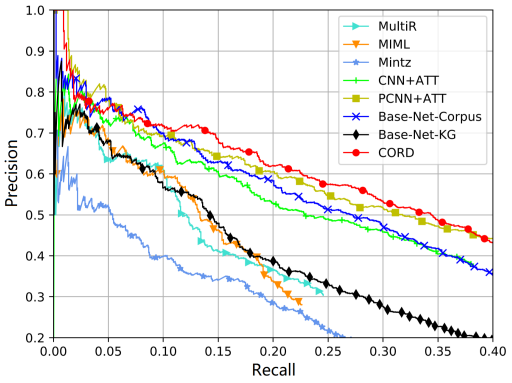

Figure 4: P-R Curve Comparison with Mintz, MIML, MultiR, CNN+ATT and PCNN+ATT.

weakness of the KG-Net might be caused by the sparsity of entity-sequence space (the dataset scales down after word-to-entity mapping), and we can enhance it by exploring other information such as relational paths (Lin et al., 2015; Zeng et al., 2017); (2) Corpus-Net achieves comparable results with *CNN+ATT* and *PCNN+ATT*, which reveals the effectiveness of Corpus-Net that could be the backbone of the CORD framework; (3) The CORD outperforms other methods on most recall area, demonstrating the effectiveness of our methods. Note that the CORD framework is significantly superior to the two separate trained base networks, especially in the rightmost area. This shows the cooperative module takes advantages of two base networks effectively and achieves better generalization. Also note that in the right-side area, CORD is still robust although the separate trained KG-Net is weak, which verifies the effectiveness of the CORD with different strengths of the base networks.

3.3 Performance of the KG Network

To evaluate the effect of incorporating KG information, we first compare P-R curve for different KG-Net setups, then we explore the benefit of using external logic rules and make a case study.

Comparison for Different Setups We experiment three KG-Nets without rule knowledge, using DeepWalk, TransE and their concatenation as entity embedding respectively. Based on whichever yields the best results, we experiment with rule knowledge and report results in Figure 5.

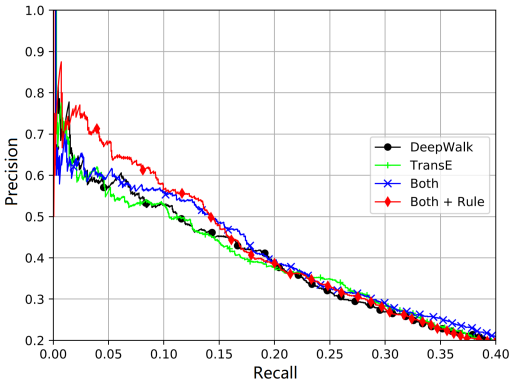

Figure 5: P-R Curve Comparison with Different Setup for the Base KG-Net.

From Figure 5, we can observe that: (1) The results of DeepWalk and TransE are slightly different and their concatenation improves, verifying the extendibility with different kinds of KG information embeddings; (2) After incorporating logic rules, the result is improved significantly as shown in left recall area, indicating that type-constraints helps to capture certain patterns more precisely.

Robustness on Long Tail Situation Efforts based on bag-level denoising such as sentence attention are liable to failure because of the long-tail situation in real-life datasets. For example, we observe that 77.63% entity pairs have only one relation instance in NYT10. We expect that supplementary external knowledge (logic rules) can enhance the robustness of this situation. We evaluate P@N of the KG-Net without rules $\mathbf{p}(\mathbf{S^e})$ and compare the rule-projected $\mathbf{p'}(\mathbf{S^e})$ on two kind of sentence amounts setup in

Table 1, where (≥ 1) means whole test dataset and (> 1) means filtering entity pairs which have only one instance.

	#Entity Pair Sentence	P@100 (%)	P@200 (%)	P@300 (%)	Average (%)
KG-Net	>1	60.2	54.0	46.4	53.5
	≥ 1	**60.3(+0.1)**	**51.5(-2.5)**	**46.0(-0.4)**	**52.6(-0.9)**
KG-Net + Rule	>1	69.2	57.0	48.6	58.3
	≥ 1	**73.5(+4.3)**	**60.4(+3.4)**	**51.7(+3.1)**	**61.9(+3.6)**

Table 1: P@N for Long Tail Situation.

From Table 1, we can observe that the KG-Net gets lower precisions (average reduces 0.9%) on the whole test data comparing with the filtered data. In contrast, the KG-Net with rules gets higher precisions on the whole test data because it can deal with noisy instances effectively in sentence-level and hence be more robust on long tail situation.

Case Study Here we pick an example instance in Table 2 to illustrate the effect of type rules. The KG-Net predicts wrong relation *place of death* probably because the appearance of entity *Joe Williams*. In contrast, it can predict correctly with the help of relation-specific type constraints.

	Predicted Relation	Relation-Specific Type
KG-Net	/people/deceased_person/place_of_death	/people/deceased_person, /location/location
KG-Net+ Rule	/location/location/contains	/location/location, /location/location
text	In ***Suffolk County***, ***Fire Island*** suffered the most damage, according to **Joe Williams**, commissioner of the county 's office of emergency management.	

Table 2: Effect of Type Constraint Rules. Bold indicates entity, italic indicates targeted entity pair.

3.4 Performance of the Cooperative Module

To investigate the effectiveness of the cooperative module, we compare the adaptive imitation rate with the fixed, the dynamic ensemble strategy with the static, and then perform a thorough case study.

Adaptive vs Fixed Imitation Rate We find that the adaptive imitation is crucial for the effective training of the CORD. To demonstrate this, we deploy some fixed imitation rate setups comparing the adaptive imitation rate. We set imitation rate (π_c, π_k) as $\{(0.5, 0.5), (0.6, 0.4), (0.4, 0.6)\}$, and report the loss curves of the dynamic, and only $(0.5, 0.5)$ for the fixed in Figure 6 because the other two have similar results.

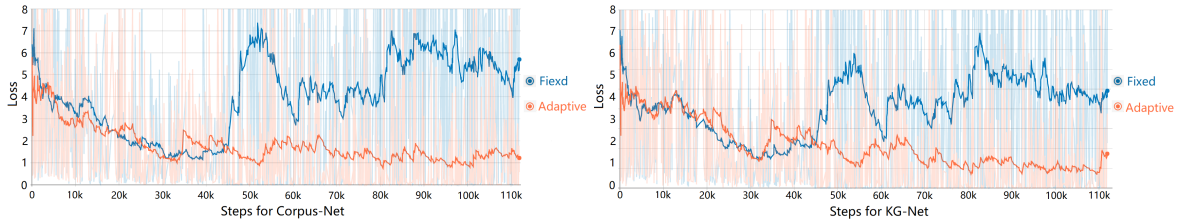

Figure 6: Loss for Adaptive and Fixed Imitation.

Figure 6 shows the remarkable difference between the fixed and the adaptive imitation, where the loss of the adaptive setting reduce gradually while the fixed fluctuates wildly. The waves of fixed KG-Net and fixed Corpus-Net are similar, meanwhile they mislead each other and the gradients are dominated by hard examples, resulting in the non-convergence. Conversely, the adaptive networks are trained effectively because the data speak for itself by providing loss values as clues to reveal the difficulty of predicting current instance. Note that the base networks are pre-trained independently and both descend gradually within the first 10k steps.

Dynamic vs Static Ensemble We also compare the dynamic ensemble strategy with the static and report their P@N results within identical base networks in Table 3. It shows that the dynamic outperforms the static and leverages the two base networks better. Note that the improvements decrease as recall

	P@100 (%)	P@200 (%)	P@300 (%)	Average (%)
Base-KG	64.0	55.2	49.3	56.2
Base-Corpus	77.3	72.0	66.5	73.9
Static	80.7	73.1	64.0	72.8
Dynamic	**84.5**	**76.0**	**66.7**	**76.1**

Table 3: P@N for Dynamic and Static Ensemble.

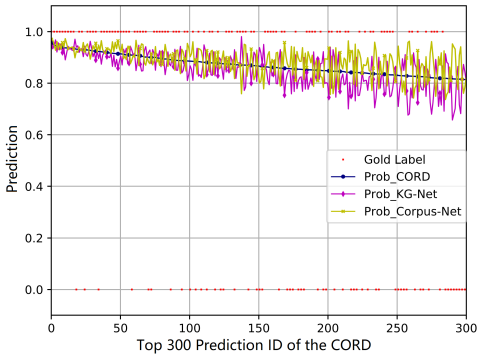

Figure 7: Top300 Predictions of the CORD.

increases, and the degree of decrease for the dynamic is less than the static, demonstrating the dynamic is more robust than the static.

Case Study To gain further insight about how the CORD works, we plot the top 300 predictions of the CORD in descending order, and compare it with its two base networks in Figure 7.

Figure 7 presents the predictions of KG-Net and Corpus-Net which go up and down around the CORD curve. The prediction differences between KG-Net and Corpus-Net are nonuniformly time-varying, indicating that some hard instances for KG-Net are easy to classify by Corpus-Net and vice versa. This supports again our view of employing adaptive imitation and dynamic ensemble.

To further demonstrate the different advantages of KG-Net and Corpus-Net, we choose two points which are predicted correctly and have significantly different values from Figure 7. The result is showed in Table 4, where prediction ID 137 and ID 257 have the different dominating network, KG-Net and Corpus-Net respectively.

ID in Figure 7	Prob-KG	Prob-Corpus	Relation	Text
137	0.9809	0.7591	location contains	when **Julian Resuello**, the mayor of ***San Carlos City*** in the northern ***Philippines***, was killed by gunmen at a campaign rally on April 28, his brother quickly stepped into his shoes.
257	0.6763	0.9583	place of death	***Ernst Haefliger***, a **Swiss** tenor who was most renowned as an interpreter of **German** art song and oratorio roles, died on Saturday in ***Davos***, **Switzerland**.

Table 4: Hard Example of KG and Corpus Net. Bold indicates entity, italic indicates targeted entity pair.

From Table 4, we can see that different networks contribute each other from the view of semantic: (1) KG-Net predicts relation *location contains* more accurately and Corpus-Net may fail if the wording of the sentence doesn't clearly state the relation between two entities. With the help of position and three entity embeddings, KG-Net can capture the relation-dependent features better from the view of graph structure. In contrast, Corpus-Net might be confused by the expression "the mayor of ..." and the uncorrelated latter part, "was killed by ..."; (2) Corpus-Net predicts relation *place of death* more accurately and KG-Net may fail if two entities are already known to be related by more than one relation. Here Corpus-Net provides reliable prediction because of the appearance of featured expression "died on ..." in the last sentence. Contrarily, KG-net may lack discriminative information and be confused by other possible relations such as *place of birth*, because the targeted person entity is followed by too many location entities.

4 Related Work

Relation extraction is one of the most important topics in NLP. Many approaches to relation extraction have been proposed, such as supervised classification (Zelenko et al., 2003; Bunescu and Mooney, 2005), bootstrapping (Carlson et al., 2010), distant supervision (Mintz et al., 2009; Krause et al., 2012; Min et al., 2013; Pershina et al., 2014; Ji et al., 2017), and generative model (Zhang et al., 2018). Among them,

distant supervision is popular as it is efficient to obtain large-scale training data automatically. However, it suffers from noisy labeling problem which severely degrades its performance.

To tackle this problem, Riedel et al. (2010; Hoffmann et al. (2011; Surdeanu et al. (2012) model distant supervision as a multi-instance learning problem under the at-least-one assumption and make it more practical. With the advances in deep learning, Zeng et al. (2015), Lin et al. (2016) and Lin et al. (2017) apply CNN and attention mechanism, Feng et al. (2017) further introduces memory network to reduce noises. Compared with these methods, the proposed framework leverages information from other sources such as KG and combine it with information from text corpus by knowledge distillation.

5 Conclusion

In this paper, we propose a novel neural relation extraction framework with bi-directional knowledge distillation to cooperatively use different information sources and alleviate the noisy label problem in distantly supervised relation extraction. Extensive experiments show that our framework can effectively model relation patterns between text corpus and KG information, and achieve the state-of-the-art results.

Acknowledgements

We thank anonymous reviewers for their helpful comments. This work was financially supported by the National Natural Science Foundation of China (No.61602013), and the Shenzhen Science and Technology Innovation Committee (Grant No. JCYJ20170412151008290 and JCYJ20170818091546869).

References

[Abujabal et al.2017] Abdalghani Abujabal, Mohamed Yahya, Mirek Riedewald, and Gerhard Weikum. 2017. Automated template generation for question answering over knowledge graphs. In *WWW*, pages 1191–1200.

[Bach et al.2017] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss Markov random fields and probabilistic soft logic. *JMLR*, 18(109):1–67.

[Bordes et al.2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

[Bordes et al.2015] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. In *NIPS*.

[Bunescu and Mooney2005] Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *HLT/EMNLP*, pages 724–731.

[Carlson et al.2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3.

[Cho et al.2014] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

[Feng et al.2017] Xiaocheng Feng, Jiang Guo, Bing Qin, Ting Liu, and Yongjie Liu. 2017. Effective deep memory networks for distant supervised relation extraction. In *IJCAI*, pages 19–25.

[Hoffmann et al.2011] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*, pages 541–550.

[Hu et al.2016] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. *ACL*.

[Ji et al.2017] Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *AAAI*, pages 3060–3066.

[Kadry and Dietz2017] Amina Kadry and Laura Dietz. 2017. Open relation extraction for support passage retrieval: Merit and open issues. In *SIGIR*, pages 1149–1152.

[Kingma and Ba2014] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Krause et al.2012] Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. In *ISWC*, pages 263–278.

[Lin et al.2015] Yankai Lin, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015. Modeling relation paths for representation learning of knowledge bases. In *EMNLP*, pages 705–714.

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

[Lin et al.2017] Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *ACL*, volume 1, pages 34–43.

[Luo et al.2017] Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *ACL*, pages 430–439.

[Min et al.2013] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *NAACL HLT*, pages 777–782.

[Mintz et al.2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*, pages 1003–1011.

[Perozzi et al.2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*, pages 701–710.

[Pershina et al.2014] Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *ACL*, volume 2, pages 732–738.

[Riedel et al.2010] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML PKDD*, pages 148–163.

[Shin et al.2015] Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. 2015. Incremental knowledge base construction using deepdive. *VLDB*, 8(11):1310–1321.

[Surdeanu et al.2012] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *EMNLP*, pages 455–465.

[Zelenko et al.2003] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *JMLR*, 3(Feb):1083–1106.

[Zeng et al.2014] 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.

[Zeng et al.2015] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762.

[Zeng et al.2017] Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Incorporating relation paths in neural relation extraction. In *EMNLP*, pages 1768–1777.

[Zhang et al.2018] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S. Yu. 2018. On the generative discovery of structured medical knowledge. In *SIGKDD*.