Meta-Learning for Neural Relation Classification with Distant Supervision

Zhenzhen Li
National University of Defense Technology
Changsha, China
lizhenzhen14@nudt.edu.cn

Jian-Yun Nie
University of Montreal
Montreal, Canada
nie@iro.umontreal.ca

Benyou Wang
University of Padua
Padova, Italy
wang@dei.unipd.it

Pan Du
University of Montreal
Montreal, Canada
pandu@iro.umontreal.ca

Yuhan Zhang
Beihang University
Beijing, China
yh_zhang@buaa.edu.cn

Lixin Zou
Baidu Inc.
Beijing, China
zoulixin15@gmail.com

Dongsheng Li
National University of Defense Technology
Changsha, China
dsli@nudt.edu.cn

ABSTRACT

Distant supervision provides a means to create a large number of weakly labeled data at low cost for relation classification. However, the resulting labeled instances are very noisy, containing data with wrong labels. Many approaches have been proposed to select a subset of reliable instances for neural model training, but they still suffer from noisy labeling problem or underutilization of the weakly-labeled data. To better select more reliable training instances, we introduce a small amount of manually labeled data as reference to guide the selection process. In this paper, we propose a meta-learning based approach, which learns to reweight noisy training data under the guidance of reference data. As the clean reference data is usually very small, we propose to augment it by dynamically distilling the most reliable elite instances from the noisy data. Experiments on several datasets demonstrate that the reference data can effectively guide the selection of training data, and our augmented approach consistently improves the performance of relation classification comparing to the existing state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Neural networks**.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412039

KEYWORDS

Distant Supervision; Relation Classification; Meta-Learning; Instance reweighting

ACM Reference Format:
Zhenzhen Li, Jian-Yun Nie, Benyou Wang, Pan Du, Yuhan Zhang, Lixin Zou, and Dongsheng Li. 2020. Meta-Learning for Neural Relation Classification with Distant Supervision. In *The 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3412039

1 INTRODUCTION

Relation classification (RC) aims to categorize the semantic relation between two entities in a sentence into a set of predefined relation types. The task is useful in various downstream applications, such as information retrieval [33], question answering [36] and knowledge graph completion [42]. Most recent methods tackle this problem by training neural classification models in a supervised way [31, 40, 42, 43], demanding a large amount of labeled training data which is difficult to acquire in practice. To alleviate the problem, distant supervision (DS) was proposed to generate abundant weakly labeled data [17, 32] by aligning facts in a knowledge base (KB) and unlabeled sentences: If a fact in the KB states a relation between two entities, then every sentence containing the same pair of entities is labeled with that relation. This process inevitably introduces many wrong labels when the sentences do not express the labeled relation. As shown in Fig. 1, if *BornIn* and *EmployedBy* are the relations between the entity pair (Barack Obama, United States) in the KB, then every sentence mentioning the same entities will be assigned both relation labels. This process generates many wrong labels (0 in Fig. 1). Training a classifier with such highly noisy labeled data will strongly limit the effectiveness of the classifier.

To alleviate the impact of wrong labels on the classifier, many existing studies [3, 7, 8, 11, 13, 15, 23, 24, 40] try to select a subset

KB={(**Barack Obama,** *BornIn,* **United States**),
(**Barack Obama,** *EmployedBy* , **United States**)}

Sentence	Assigned labels	Sentence-level evaluation
Barack Obama is the 44th President of the **United States**.	*EmployedBy*	*1*
Barack Obama is the 44th President of the **United States**.	*BornIn*	*0*
Obama ran for the **United States** Senate in 2004.	*EmployedBy*	*0*
Obama ran for the **United States** Senate in 2004.	*BornIn*	*0*

Figure 1: Examples of sentence-level evaluation for DS labeled data, which evaluates each instance individually. "1" and "0" denote the confidence score about the labelling quality.

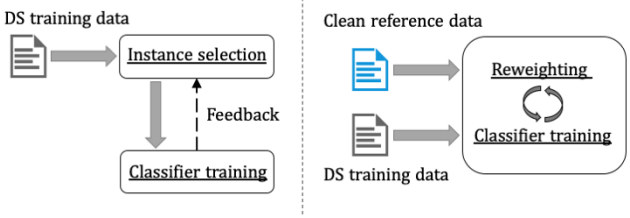

Figure 2: Bootstrapping process from noisy data (left) vs. our meta-learning approach guided by reference data (right).

of reliable instances for training. Some approaches [7, 13, 15, 40] adopt multi-instance learning, which relaxes the relation label for each sentence to a bag of sentences mentioning the same entity pair and assumes that at least one sentence expresses the relation. These approaches are often used in relation extraction, i.e. to identify possible relations between a pair of entities. However, for our task of sentence-level relation classification (i.e. to judge if a given sentence expresses a specific relation), they turn out to be ineffective [3, 8], because we have to judge on each of the instances rather than a bag. From this perspective, sentence-level relation classification is fine-grained and thus more difficult than relation extraction. Some recent approaches [3, 5, 8, 23, 24, 41] perform sentence-level evaluation. They evaluate the labeling quality of each instance and select those deemed reliable for model training. These approaches leverage the large amount of distantly labeled instances to emerge strong supervision signals during the iterative training process. As illustrated in the left part of Fig. 2, they usually resort to reinforcement learning or adversarial learning to train an instance selection model by receiving feedback from the classifier or a manually crafted reward function. The basic assumption is that most training instances are true, and thus the iterative learning process will help select conïňĄdent instances whose predicted label by the learned classifier is consistent with its DS-generated label. We put these approaches in the category of *bootstrapping* approaches, i.e. an instance selection model emerges itself from the noisy data gradually. Even though the bootstrapping process can learn a good instance selection model, it may be easily trapped by some common wrong instances. In addition, the instance selector may fail to select some true instances that are different from the frequent instances, thus underutilize the useful training data. Existing approaches either retain too many noisy training instances or miss a lot of effective training data, limiting the capability of the resulting classifier.

The above problem stems from the fact that the selector is trained without further supervision signals. If we have some manually labeled instances that show good examples to express a specific relation in natural language, then the selection model can be guided by them. So, our idea in this paper is to use a small set of manually annotated samples as "reference data" to guide the instance selection. This corresponds to a new setting of DS for relation classification, where a small set of reference data is available in addition to a large amount of noisy DS data. This setting is realistic - in practice, it is not difficult to label a small set of data if this turns out to be beneficial, which will be shown in our experiments.

In this paper, we focus on effective utilization of the small set of reference data for relation classification using DS. We propose a meta-learning framework, in which the reference data is used to guide instance weighting at the meta-learning level. Then the weighted instances are used to train a classifier. We use a specific meta-learning method similar to the online reweighting algorithm [26, 28], which performs a meta gradient descent step to adjust the instance weights in order to minimize the classification loss on the clean reference data. This method is used because it can more effectively leverage clean reference data than existing reinforcement learning or adversarial learning approaches [3, 5, 23, 24, 41].

As we assume that only a small set of clean data is available, we may have the mode collapse issues: it may over-boost the noisy samples that agree with the reference samples, leading the classifier to collapse into the existing mode and thus a poor generalization capability. To relieve this issue, we propose to augment the clean reference data by a set of noisy instances deemed confident. We call them "elite instances". At each training iteration, the elite instances are used together with the clean data to guide a *reweighting* process to assign weights to noisy instances, based on which training instances are selected. The process can combines the strengths of bootstrapping from noisy data and guidance by reference data. The whole training process iterates between reweighting and classifier training as shown in the right side of Fig. 2.

The contributions in this work include:

- We propose an effective way to select noisy data for training, guided by a small set of reference data.
- We adopt the meta-learning mechanism for DS relation classification for the first time, and show its high effectiveness compared to other alternative methods.
- Our approach combines the strengths of supervision and bootstrapping to create a larger set of reference data for meta-learning.

2 METHODOLOGY

Problem Definition The relation classification problem can be formulated as follows: Let $\mathbb{D}_{\text{train}}$ be a set of sentence-relation pairs $\{(x_i, r_i)\}_{i=1}^{N}$, where r_i is a relation label created by distant supervision. The goal is to train a relation classification model Φ parameterized by θ that can predict the relation of a new sentence x_j from test data $\mathbb{D}_{\text{test}}$, i.e., compute the probability $P_\Phi(r_j|x_j)$. Formally, the task of learning under distant supervision is to minimize the empirical risk on training data as follows:

$$\begin{aligned}\theta^* &= \operatorname*{argmin}_{\theta} \left(\mathbb{E}_{(x,r)\sim p_{\mathbb{D}_{\text{train}}}} (\ell(\Phi_\theta(x), r))\right) \\ &= \operatorname*{argmin}_{\theta} \left(\sum_{(x,r)} p_{\mathbb{D}_{\text{train}}}(x,r)\ell(\Phi_\theta(x), r)\right) \\ &\approx \operatorname*{argmin}_{\theta} \left(\sum_{i=1}^{N} w_i \ell(\Phi_\theta(x_i), r_i)\right).\end{aligned} \tag{1}$$

In the last step, the training data is used with their weights $\{w_i\}_{i=1}^N$ representing the labeling quality.

The selection of good instances is a key problem in DS learning. Different from the previous approaches [3, 5, 23, 24, 41], we assume that we have a small set of clean reference data $\mathbb{D}_{\text{ref}}$ that can be used to guide the selection or weighting process.

2.1 Overview of the Proposed Approach

We propose to adopt a weighting schema of training samples to dynamically reweight noisy instances in $\mathbb{D}_{\text{train}}$ under the guidance of $\mathbb{D}_{\text{ref}}$. We achieve this goal by proposing an approach based on meta-learning, which aims at minimizing the meta-objective:

$$w^* = \operatorname*{argmin}_{w} \text{MetaObjective}(\Phi_\theta, \mathbb{D}_{\text{ref}}) \tag{2}$$

The principle of our meta-learning is as follows: we want the classifier trained with the weighted instances to minimize the loss on the reference data. Note that Φ_θ is a function of w since w affects the optimization of θ as shown in Eq. 1. We will provide details about this instance reweighting (meta-reweighting) process later.

As we mentioned earlier, we assume that we only have a small set of reference data. A typical consequence is mode collapse, i.e., the selected training instances (assigned high weights) become similar to the reference data, leading to a model with narrow coverage. To cope with this challenge, we increase the coverage of the reference data by augmenting it with some highly reliable instances from the noisy set, which we call "elite data" or "elite instances". The elite data is expected to represent some strong patterns of a relation among the noisy data, which may not be covered by the clean reference data.

The framework of our proposed approach is shown in the Fig. 3, consisting of iterations of **robust classifier training** based on meta-reweighting and **elite instance selection** to augment reference data (to extend reliable supervision signals). Details are given in the following subsections.

2.2 Robust Classifier Training by Meta Learning

Generally, the goal of machine learning is to find a model Φ with the best parameter θ to minimize the empirical risk as in Eq. 1. In the DS setting, this means to weight the training instances (w) to train a better model Φ. We propose to leverage the small set $\mathbb{D}_{\text{ref}}$ for this purpose using meta-learning based sample reweighting algorithm [26, 28]. As shown in the Fig. 4, our meta-learning algorithm involves two optimization process: the outer loop for instance reweighting and the inner loop for classifier training. As

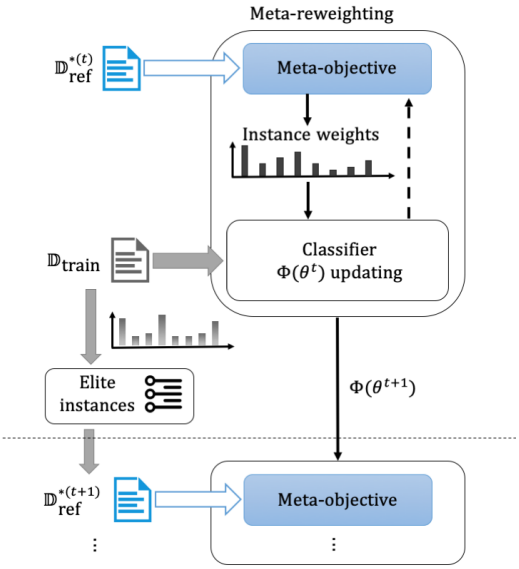

Figure 3: Framework overview. At each training epoch t, we learn to reweight noisy $\mathbb{D}_{\text{train}}$ according to a meta-objective with respect to the augmented reference data. The classifier $\Phi(\theta^t)$ is updated with the weighted instances. Afterwards, we select a set of elite instances to augment the clean reference data to $\mathbb{D}_{\text{ref}}^{*(t+1)}$ for the next epoch.

the outer loop is optimized for better inner loop, the classifier training is embedded in the instance reweighting process. That is, at each iteration,we start with the classifier training with weighted training data, and then optimize w based on the updated classifier with respect to the meta-objective (on reference data). Once w is optimized, we use it to update the model Φ in the classifier training. The whole training process iterates between the instance reweighting and classifier training.

To increase the training efficiency, we adopt the online reweighting strategy [26], which dynamically learns the instances weights for a mini-batch of training data by a single optimization step. In the **instance reweighting** phase, given a mini batch of training instances $\mathbb{D}_{\text{mtrain}} \subset \mathbb{D}_{\text{train}}$ and weight vector w (which may be initialized by perturbing [10]), we define the weighted training loss as follows:

$$\mathcal{L}_{\text{mtrain}}(\theta(w)) = \sum_{(x_i,r_i)\in\mathbb{D}_{\text{mtrain}}} w_i \ell(\Phi_\theta(x_i), r_i). \tag{3}$$

We update $\theta(w)$ to a temporary version $\hat{\theta}(w)$ as Eq. 4, which is only used to optimize instance weights:

$$\hat{\theta}(w) = \theta - \eta_t \nabla_\theta (\mathcal{L}_{\text{mtrain}}(\theta(w)) \tag{4}$$

Based on such a fixed classifier $\Phi_{\hat{\theta}(w)}$, the meta-objective is defined as the loss on the reference data:

$$\begin{aligned}\text{MetaObjective}\left(\hat{\theta}(w)\right) &\triangleq \mathcal{L}_{\text{ref}}\left(\hat{\theta}(w)\right) \\ &= \sum_{(x_i,r_i)\in\mathbb{D}_{\text{ref}}} \ell(\Phi_{\hat{\theta}(w)}(x_i), r_i).\end{aligned} \tag{5}$$

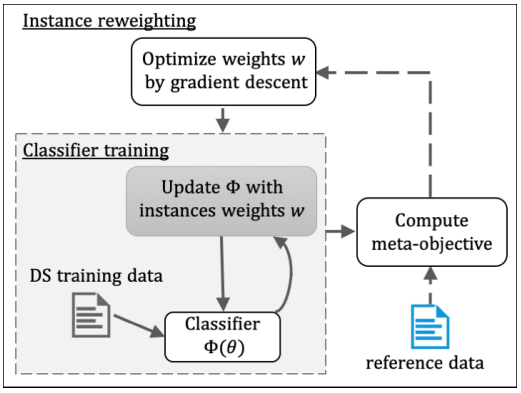

Figure 4: Schematic diagram of meta-learning algorithm for robust classifier training, which iterates between instance reweighting and classifier training. The instance reweighting process contains extra classifier training process.

By applying gradient descent to minimize such meta-objective, we can optimize w through second order derivatives. Thus, we get our new weight vector as:

$$w^* = \underset{w:w\geq 0}{\operatorname{argmin}} \text{MetaObjective}\left(\hat{\theta}(w)\right) \tag{6}$$

With the learned instance weights w^*, we can update the classifier according to Eq. 7. This phase is the actual **classifier training** phase:

$$\theta' = \theta - \eta_t \nabla_\theta \left(\sum_{(x_i,r_i)\in\mathbb{D}_{\text{mtrain}}} w_i^* \ell(\Phi_\theta(x_i), r_i) \right). \tag{7}$$

The updated θ' are used as the parameters of Φ in the next iteration.

It is worth noting that w is optimized to make the classifier perform better on $\mathbb{D}_{\text{ref}}$ as Eq. 6 expresses. Further analysis [26, 28] shows that the training instances whose gradient directions are similar to the gradient direction of $\mathbb{D}_{\text{ref}}$ will be assigned with high weights, otherwise low weights. With such bilevel optimization process, we could maximize the role of reference data for denoising.

2.3 Enhancing Reference Data by Elite Instances

The small amount of clean reference data could provide limited guidance when participating in the meta-objective, resulting in poor generalization capability. To overcome this issue, we augment the reference data by distilled elite instances from the noisy data. The elite instances are those that are attributed to the highest classification sores, thus are highly reliable. The elite instances are "elected" from the noisy data by the classifier they trained. They are thus the most representative likely-true instances among the noisy data. The elite instances are intended to provide additional reliable supervision signals to the reference data, as we explained earlier.

2.3.1 Enhancement strategy. We propose to augment the reference set dynamically, i.e., we evaluate training data at each epoch and select top-scored instances to expand the original reference set $\mathbb{D}_{\text{ref}}$ to $\mathbb{D}^*_{\text{ref}}$. Since the top-scored instances vary for each training epoch, we are able to leverage diverse elite instances to mitigate the "mode collapse issue".

We propose two strategies to compute the confidence score of labeling quality based on respectively the instance weighting (sw) and classification probability (sp). In the first strategy, we use the normalized instance weight determined with respect to the clean reference data (the same as in meta-learning) as the confidence score, defined as $sw_i = \frac{w_i}{\sum w_j}$. Since the instances assigned with high weights have gradient directions similar to the reference data [26, 28], sw may help select instances consistent with the reference data.

The second strategy uses the prediction probability by the classifier, defined by Eq. 8.

$$sp_i = \frac{P_\Phi(r_i|x_i)}{\sum P_\Phi(r_j|x_j)} \tag{8}$$

As the classifier is trained with a large amount of weighted noisy data, the instances whose labels are scored high by the classifier are more likely to be consistent with the common patterns emerged from boosted training data. The advantage of using sp is that its confidence score becomes more stable as the classifier training converges, while the confidence score by sw is learnt online and fluctuates across training epochs. We further examine this in Section 2.3.2.

With the confidence score s_i (either sw_i or sp_i), we rank and select instances from top to bottom. We select the top-scored instances for each type of relation, proportionally to the number of that relation in $\mathbb{D}_{\text{ref}}$. Note that we follow the same relation distribution as $\mathbb{D}_{\text{ref}}$ in order to avoid introducing unexpected bias from unbalanced relation distribution of $\mathbb{D}_{\text{train}}$.

Formally, let $\mathbb{RE}$ be the set of target relation types excluding the not-the-target relation type[1]; n_r be the number of selected instances for relation r, set as $k\#\{(x_i,r_i)|(x_i,r_i)\in\mathbb{D}_{\text{ref}}, r_i = r\}$ [2], where k is the expansion ratio. Then the expanded reference data is as follows:

$$\mathbb{D}^{(t)}_{\text{exp}} = \bigcup_{r\in\mathbb{RE}} \left\{(x_i,r)|s_i \geq \tau(n_r), (x_i,r)\in\mathbb{D}_{\text{train}}\right\}, \tag{9}$$

where $\tau(n_r)$ is the minimum confidence score of top n_r.

Finally, the augmented reference data is the union of the original one and the expanded one, i.e., $\mathbb{D}^{*(t)}_{\text{ref}} = \mathbb{D}_{\text{ref}} \cup \mathbb{D}^{(t)}_{\text{exp}}$. The meta-objective in Eq. 5 is enhanced as:

$$\text{MetaObjective}\left(\hat{\theta}(w)\right) = \mathcal{L}_{\text{ref}}\left(\hat{\theta}(w)\right) + \beta\mathcal{L}_{\text{exp}}\left(\hat{\theta}(w)\right), \tag{10}$$

where β is a factor to control the influence of expanded reference data for meta-objective.

2.3.2 Robust exploitation phase. The aforementioned strategy enhances instance reweighting, but may still suffer from wrong labels in the expanded reference set. To further improve the quality of selected elite instances, we propose an additional exploitation strategy based on the assumption that the more times an instance has been selected as elite instance in different epochs, the more likely it is true. Since confidence scores among different training epochs are incomparable, we design an accumulate score from the ranking

[1] We do not select instances of not-the-target type, i.e. the negative training instances, since they are dominant in the $\mathbb{D}_{\text{train}}$, $\mathbb{D}_{\text{ref}}$. Selecting such instances has been empirically found useless in our preliminary experiments.
[2] $\#\{.\}$ is the function of counting the number of a set.

order to record the historical selection information. The accumulative score sa_i for each training instance (x_i, r_i) at training epoch t is defined as follows:

$$sa_i^{(t)} = \begin{cases} 0, & \text{if } t = 0 \\ \gamma sa_i^{(t-1)} + sr_i^{(t)}, & \text{otherwise} \end{cases} \quad (11)$$

where γ is a parameter used to decay the influence of distant historical evaluations, and $sr_i^{(t)}$ is the ranking score at epoch t:

$$sr_i = \frac{1}{1 + e^{\text{Idx}_i - n_r}}. \quad (12)$$

where Idx_i is its rank according to sw_i or sp_i from top to bottom.

We apply the exploitation strategy only after a certain number of training epochs (in our experiments, half of the total training epochs) in order to allow for more diversified selection at the beginning (*aka* exploration).

2.4 Algorithm

We design an iterative algorithm for model training, as shown in Algorithm 1. At the training epoch t, we compute the current confidence score sp_i (line 4) (which could also be sw_i) and update it to the accumulative score sa_i (line 5), which is used in exploitation phase. Then we select top-scored instances according to s_i to construct expanded reference set (line 8).

ALGORITHM 1: Our approach for robust classifier training from noisy data.

Require: Noisy training dataset $\mathbb{D}_{\text{train}}$;
Clean reference dataset $\mathbb{D}_{\text{ref}}$;
Classifier Φ parameterized by θ
Ensure: $\theta^{(L)}$

1: Pre-train Φ (with parameters $\theta^{(0)}$);
2: **for** epoch $t = 1$ to L **do**
3: **for each** (x_i, r_i) in $\mathbb{D}_{\text{train}}$ **do**
4: Compute confidence score sp_i with $\Phi(\theta^{(t)})$ as Eq. 8.
5: Accumulate ranking score $sa_i^{(t)}$ as Eq. 11
6: Set final score s_i as $sa_i^{(t)}$ if in exploitation phase otherwise sp_i (Section. 2.3.2)
7: **end for**
8: Select top-scored instances based on s_i to obtain expanded reference set $\mathbb{D}_{\text{exp}}^{(t)}$ (Section 2.3.1)
9: **for** each mini batch in $\mathbb{D}_{\text{train}}$ **do**
10: Calculate meta-objective with the loss on $\mathbb{D}_{\text{ref}}$ and $\mathbb{D}_{\text{exp}}^{(t)}$ as Eq. 10.
11: Optimize w^* based on the meta-objective as Eq. 6.
12: Use w^* to update classifier $\Phi(\theta)$ as Eq. 7
13: **end for**
14: **end for**

Then the augmented reference data is applied to the online reweighting algorithm, which updates w^* and θ using mini-batches of training data (lines 11 - 12), similar to the existing meta-reweighting approach [26, 28].

3 EXPERIMENTAL SETUP

3.1 Datasets

We evaluate our approach on two widely-used DS datasets for relation extraction/classification: Wiki-KBP and NYT. To evaluate our models precisely, we construct a new version of Wiki-KBP and adopt a manually labeled test set for NYT (see explanations below). The statistics about the datasets are presented in Table 1.

Table 1: Statistics of the datasets used in our experiments.

Datasets		#Relation types*	#Instances	#Positive instances
Wiki-KBP	Training	6	151,091	38,922
	Test	6	4,168	1,075
NYT	Training	53	522,611	136,947
	Test	17	2,040	684

*Including the Not-Target-Type for negative instances.

Wiki-KBP The training dataset was constructed by aligning Freebase facts with English Wikipedia corpus [14]. The commonly-used test set was manually annotated from the 2013 KBP corpus [2]. However, the relation types in the test set do not always match with those in the training set - some test relation types have no or only one training instance. We thus removed these relation types.

The resulting test set contains a very small number of positive instances (165) and their proportion in the test set is only about 0.08. To have more labeled sentences of target relations for evaluating classifiers, we increase the number of positive instances in the test set by merging another manually annotated dataset - TACRED [42], whose sentences are also derived from the KBP corpus.

Specifically, we removed the relation with only one instance in the training set (per:countries_of_residence), and obtained the modified training set with 5 actual relations - per:country_of_death, per:country_of_birth, per:children, per:parents, per:religion, and a special relation not-target-type (None) for negative instances. Then, we merged the positive instances of these 5 types from the original test set [3] and the TACRED dataset [4]. To make sure that our test set has a similar proportion of positive instances as the training set, we randomly sampled 3093 negative instances ("no_relation") from the TACRED dataset. The resulting training and test sets are as shown in Table 1.

NYT This dataset was generated by aligning news corpus from New York Times (NYT)[5] and relation facts in Freebase [27]. Most previous work used held-out evaluation, where the training set and test set were heuristically annotated based on disjoint sets of the freebase facts. However, the DS-generated labels could be wrong and the two sets are noisy. To evaluate the relation classification at sentence level, we used the original training set and a manually annotated test set provided by Jiang et al. [9].

The **clean reference data** is constructed as follows: For Wiki-KBP, to better evaluate our models, we split the original test dataset into a validation set and a test set. We fixed the test set and randomly sampled **100 instances** from the clean validation set as the reference data. For NYT, since the manually labeled data is very

[3] https://github.com/shanzhenren/CoType
[4] https://nlp.stanford.edu/projects/tacred/
[5] http://iesl.cs.umass.edu/riedel/ecml/

small, we sampled 10% of the training data as validation set and left the rest as the actual training set. To keep the original setting in which no entity pair overlaps between the training and test sets, we randomly sampled one fifth entity pairs from the test dataset and extracted the corresponding sentences (**about 400 instances**) as the reference data. The remaining test data was used as actual test set.

Note that we set the minimal size of reference data by making sure that there is at least one instance for each relation in the reference set. We randomly sampled five sets of reference data and fixed them to conduct our experiments. All the reported average results for baselines and analysis experiments are based on the same reference sets.

3.2 Compared models

3.2.1 Our model. Based on the model-agnostic meta-learning algorithm, our approach can be applied to any neural network architectures for DS relation classification.

In this work, we adopt the Piecewise Convolutional Neural Network (PCNN) [39] as base model to compare our approach with other denoising approaches. This model has been widely used for relation classification with DS [7, 13, 34, 39] and been proven to perform better than CNN models [9]. The PCNN model contains an input layer which concatenates the word embedding and the position embedding, a convolutional layer, a piecewise max pooling layer, and a softmax layer. More details can be found in [39].

Based on the general meta-learning mechanism described in section 2.2, we build our approach on top of learning to reweight examples (L2RW) [26] algorithm. It regards instances weights as a meta-parameter vector implicitly learned and does not require extra hyper-parameter tuning. Instance weights can also be learned by a parameterized module within the meta-learning framework, such as a multilayer perceptron network (Meta-Weight-Net) [28]. We will not examine it in this paper.

We measure the effectiveness of two strategies for elite instance selection, i.e. *sw* (by online-learnt instance weights) and *sp* (by prediction probability from the classifier). Then we use *sp* as the default to select elite instances for detailed analysis in section 4.2.

3.2.2 Baselines. We compare our approach with previous representative and state-of-the-art instance selection approaches in DS relation classification, including:

- **PCNN+ATT** [13] is a classical bag-level instance weighting approach. It assigns attention weights to each instance within a bag according to their relevance to the bag label, and thus down-weights the relatively noisy instances.
- **PCNN+RL** [24] adopts reinforcement learning to generate the false-positive indicator to recognize false positives, and then redistributes them to the negative set to obtain a new cleaned dataset.
- **PCNN+DSGAN** [23] adopts adversarial learning to train a generator to recognize true positive instances, and then redistributes the remaining false positives to the negative set to obtain a new cleaned dataset.
- **ARNOR** [8] is previous state-of-the-art model for DS relation classification. It starts the model training with reliable instances selected by a set of frequent relation patterns, and then adds patterns by bootstrapping.

We apply the reference data to enhance existing models as follows: 1) for RL [24] and DSGAN [23], we compute reward on reference data with the same reward function and average it with previous rewards to obtain final feedback for instance selection; 2) for ARNOR [8], we add the relation patterns extracted from the reference set to their pattern collections.

We compare with the following baseline models that use extra clean data:

- **PCNN+L2RW** [26] is the learning to reweight samples algorithm that resorts to a clean unbiased dataset to tackle the biased training set problem.
- **PCNN+BA** [37] is a bias adjustment approach that uses the extra clean data to help relieve the label distribution shift from DS training data and manually annotated test data. We use their best-performing BA-Fix model.

3.3 Implementation Details

We use the same word embedings for baselines, i.e., the pre-trained GloVe [21] embedding[6] for Wiki-KBP and the 50-dimensional word embedding file[7] released by Lin et al. [13] for NYT. The corresponding position embeddings are set as 30 dimensions for Wiki-KBP and 5 dimensions for NYT. The number of convolution filter for PCNN model is 230, and the filter window size is 3.

For our enhanced reweighting strategy, we select k times elite instances compared to the original reference data. k falls between 2 and 3, depending on the dataset. Parameter β is generally set as 1, and adjusted as 0.1 for *sw* during the exploitation phase. γ is set as 0.97. For model training, we adopt the SGD optimizer and set the maximum training epoch as 25. The batch size for training set is 160 and we use all the reference data for every reweighting step due to its small size. To speed up the training process, we linearly warm up the learning rate to a maximum value of 0.1 within 2 epochs, and linearly decrease it to a minimum value of 0.001 used for the last 5 epochs.

All the baselines were implemented with the source codes released by their authors except for ARNOR - ARNOR is not open-sourced, so we reimplemented it based on an open implementation[8], which achieves comparable performance on the same dataset as reported in [8]. Since ARNOR starts by selecting reliable instances based on several frequent relation patterns, when training instances are spread over many patterns (e.g. in Wiki-KBP), it fails to select enough training data. Therefore, we increased the number of patterns when applied it to our new dataset to increase its performance.

4 EVALUATION RESULTS

We use the same metrics as previous work [8, 42]: micro-averaged Precision (Pre.), Recall (Rec.) and F1-score (F1).

Table 2: Test set performance comparison of different models on the Wiki-KBP and NYT datasets. The five-time average and standard deviation of test results are reported as percentage, and the best (bold) and second best (*italic*) F1 scores are highlighted below.

Model	Wiki-KBP			NYT		
	Prec.	Rec.	F1	Prec.	Rec.	F1
PCNN$^{\triangle}$ [39]	55.39 ± 3.75	34.37 ± 4.14	42.16 ± 1.97	46.93 ± 1.68	57.79 ± 5.01	51.66 ± 1.61
PCNN [39]	56.12 ± 3.33	33.38 ± 2.17	41.75 ± 0.98	47.88 ± 1.81	57.38 ± 3.75	52.12 ± 1.41
PCNN+ATT [13]	72.65 ± 1.99	29.24 ± 1.31	41.69 ± 1.55	59.99 ± 1.86	49.79 ± 2.29	54.36 ± 0.79
PCNN+RL [24]	57.64 ± 2.41	38.79 ± 2.22	46.32 ± 1.65	48.72 ± 1.72	48.93 ± 2.11	48.78 ± 0.46
PCNN+DSGAN [23]	59.86 ± 5.65	38.54 ± 2.97	46.65 ± 1.19	47.55 ± 1.15	51.52 ± 1.30	49.44 ± 0.75
ARNOR [8]	54.83 ± 2.40	34.59 ± 2.20	42.35 ± 1.32	68.39 ± 1.37	48.42 ± 2.66	56.67 ± 2.06
PCNN+BA [37]	58.86 ± 2.92	38.71 ± 2.60	46.59 ± 2.73	45.33 ± 0.86	58.06 ± 3.24	50.70 ± 1.21
PCNN+L2RW [26]	66.22 ± 3.36	43.80 ± 4.72	52.56 ± 3.45	66.23 ± 4.17	53.05 ± 3.40	58.56 ± 2.12
OURS_sw	60.15 ± 3.58	51.43 ± 5.05	*54.56* ± 3.94*	63.65 ± 4.63	57.79 ± 4.54	*60.38* ± 2.54*
OURS_sp	67.24 ± 3.10	46.65 ± 4.13	**54.98* ± 3.00**	67.05 ± 4.55	56.59 ± 2.25	**61.26* ± 1.98**

$^{\triangle}$ represents the model trained without using clean data.

* indicates statistically significant improvements over the L2RW (i.e. Wilcoxon signed-rank test with $p < 0.05$)[9].

4.1 Overall Results

From the results shown in Table 2, the following observations can be made:

- Our approach outperforms all the baselines on F1 score and improves the F1 score of base PCNN model (line 2) by over 12% on Wiki-KBP and 8% on NYT. This demonstrates the higher effectiveness of our approach.
- L2RW and our approach outperform other instance selection approaches. This shows that meta-learning based approaches are able to make better use of the clean reference data than simply adding it to the training data (PCNN in line 2), using its relation patterns to extract training data (ARNOR), and computing reward for instance selection under existing RL or DSGAN framework. We also empirically found that fine-tuning the PCNN model with clean reference data does not improve its performance and may even decrease it when the size of clean reference data is too small.
- Both strategies of our approach improve L2RW by a statistically significant margin. This shows that our enhanced strategy for instance reweighting is beneficial. We observe that L2RW can substantially improve the precision of PCNN model on two datasets, but decrease the recall on NYT. This observation confirms our intuition that the small size of clean reference data may lead to mode collapse. In contrast, with our enhanced strategy, the recall score and F1 score are consistently increased. This indicates that the expanded reference data could enhance the reweighting process to select more reliable training data with new patterns, thus improving the generalization capability.
- The bias adjustment approach (PCNN+BA model) does not show stable improvements on two datasets compared to the base model. The F1 score of PCNN+BA is slightly decreased on NYT. We explain this by the fact that the reference set and test set in NYT are divided based on entity pairs, and thus the reference set does not provide exactly matched label distribution information. This shows that the small set of clean reference data could provide little information about label distribution, but our approach using it for denoising makes a big difference.

4.2 Analysis and Discussion

4.2.1 Understanding the impact of reference data. Our approach reweights the noisy training data under the guidance of augmented reference data during the classifier training. To understand the impact of reference data, we present the performance of classifiers that are trained under three settings: with original noisy training data (PCNN), with meta-reweighting algorithm guided by clean reference data (L2RW), with meta-reweighting algorithm guided by augmented reference data (OURS).

As shown in Fig 5, after several training epochs, PCNN model quickly stagnates, while our approach and L2RW continue to improve the performance with a large margin. This shows the impact of the meta-reweighting for learning under DS. During the exploitation phase (after epoch 12), our enhanced strategy exploits *sa* to distill elite instances and uses them to guide the reweighting process. We see steadily superior performance over L2RW. This is a demonstration of the usefulness of elite instances.

During the exploration phase (before epoch 12), our augmenting strategy selects elite instances according to the current confidence score. Fig. 5 shows that it may perform poorer than L2RW sometimes. This may be due to the fact that the selected elite instances may contain noise. As the training epoch increases, the well-trained classifier and *sa* could help distill more reliable elite instances, thus enhancing the reweighting. At the end, our final model after all the training epochs is better than the others.

4.2.2 Impact of exploitation strategy. To reduce possible false instances in elite data when using the current confidence score (i.e. *sw* or *sp*) for selection, we propose to use accumulative score (*sa*) during the exploitation phase (after epoch 12). To verify the effectiveness of the exploitation strategy, we test our approach with and without

[6]http://nlp.stanford.edu/data/glove.840B.300d.zip
[7]https://github.com/thunlp/NRE
[8]https://github.com/HeYilong0316/ARNOR
[9] We use the Wilcoxon signed-rank test to measure the paired results of the L2RW and OURS based on the same set of clean reference data, and the t-test results are computed with the scipy package.

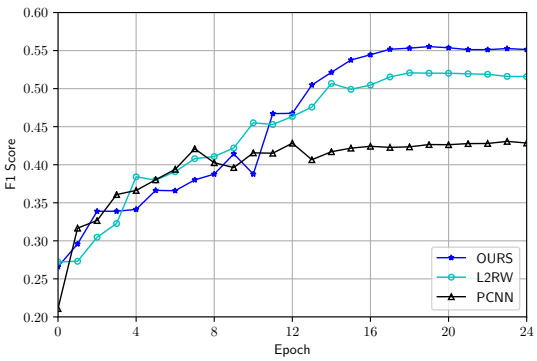

Figure 5: Test set F1 score of three models on Wiki-KBP at each epoch. The lines are drawn with average F1 scores based on five reference sets.

using the accumulative score (with *sa* vs. w/o *sa*) and compare them with the algorithm without enhancement strategy (L2RW). We apply meta-reweighting algorithms to train the PCNN model based on one set of reference data and present their smoothed convergence curves of 5-step moving averages on both the validation and test sets of Wiki-KBP.

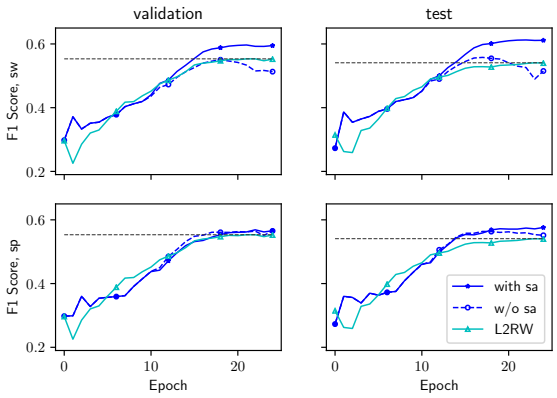

Figure 6: Smoothed convergence curves of F1 scores for models using *sa* or not on both validation set and test set of Wiki-KBP. Results for two evaluation criteria *sw* and *sp* are presented respectively.

As shown in Fig. 6, *sa* is beneficial for both evaluation criteria (*sw* or *sp*) and is especially important for *sw*. When *sa* is not used, the performance of models using *sw* decreases at the end of training process and may be even poorer than L2RW sometimes. This shows that noise elite instances may be selected into the expanded reference set. When using *sa*, more reliable elite instances are selected. The improvements are more steady. Note that when using other reference sets the improvement margin over L2RW may vary, but similar impact of the exploitation strategy is observed. The above observation shows that the exploitation strategy that takes into account the historical evaluations is beneficial, especially for the scenario where evaluation results fluctuate across epochs.

4.2.3 Size of the clean reference data. To see the impact of the size of initial clean reference data on meta-reweighting algorithm, we evaluate the performance of models when the size of clean reference data is increased from the initial 100. Following the same setting, we randomly sample the required reference set five times from the validation set of Wiki-KBP, and present the average results with L2RW and enhanced strategy in Fig. 7. We can see that when the number of clean reference data increases, both our model and L2RW can further improve the performance of the classifier. When 400 clean reference samples are available, our approach achieves 67.46% F1 score, which improves the basic PCNN model by more than 25%. However, we also see that the gap with L2RW is reduced. This suggests that our enhancement with elite data is more useful when the size of the clean reference data is relatively small. We explain this by the fact that when there is sufficient clean reference data, it becomes less critical to further extend it because it already has a quite good coverage of different relation patterns. So, our enhancement is the most beneficial with a small amount of clean reference data - the situation we target in this paper.

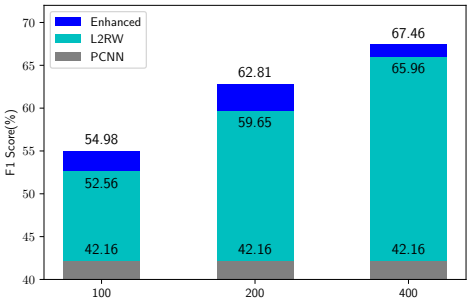

Figure 7: Test set F1 score improvements compared with PCNN model when different sizes of clean reference data are used by our approach in Wiki-KBP dataset.

4.2.4 Case study. To better understand the impact of the meta-reweighting algorithm and our expanded reference instances, we analyze the case of one particular relation. The case shown in Table 3 is an example extracted from our data for relation '/location/administrative_division/country' in NYT dataset, and Table 4 presents examples about sentence weighting and relation classification. To make our analysis easier, we only consider using similar words between two entities to match sentences, which is also regarded as the relation pattern in ARNOR [8].

Effectiveness of meta-reweighting. The relation patterns of clean reference data are all used by ARNOR to select training data with the same relation patterns, such as "entity1 , in southern entity2" in sentence 2. However, it fails to select some useful training sentences, such as sentences 6 and 8, due to their different patterns. In contrast, the reweighting algorithms (L2RW and OURS) can assign a positive weight to sentence 6 which is similar to sentence

Relation: /location/administrative_division/country		
Dataset	ID	Sentences
Clean reference	1	... a company in **Auroville**, **India**
	2	... on the school in **Beslan**, in southern **Russia**
	3	... "cities on the rise": **Riga**, **Latvia**
Expanded reference	4	... for nearby **Aleppo** in **Syria**
	5	... family's zoo in **Queensland**, in northern **Australia**

Table 3: Examples showing that our elite instances enrich patterns of clean reference data.

Relation: /location/administrative_division/country				
ID	Sentences	ARNOR	L2RW	OURS
6	... tour in **Chechnya** crashed in southern **Russia**	0	0.002	0.003
7	... in outback **New_south_wales** in **Australia**	1	0.0	0.004
8	... **Haryana** and Uttar Pradesh in northern **India** ...	0	0.0	0.003
9	... winter olympics in **Albertville** , **France**.	×	√	√
10	... a Philistine seaport at **Ashkelon** in **Israel**.	√	×	√
11	... here in rural **Bihar** state in northern **India** ...	×	×	√

Table 4: Examples about sentence weighting (the upper part) and relation classification (the lower part). We underline similar relation expressions with the reference samples.

2. This confirms that the meta-reweighting algorithms could make better use of the limited supervision signal (the clean reference data) to boost relevant training data than using the shallow patterns to select training data (ARNOR).

Effectiveness of expanded reference data. Table 3 shows that the distilled elite instances from noisy training data could provide not only similar relation expression (sentence 5 which is similar to sentence 2), but also new expression patterns such as "entity1 in entity2" in sentence 4. This shows that the expanded reference has the capability of covering more relation expression cases.

Once our distilled elite instances (e.g. sentence 4 and sentence 5) are added into the reference set, sentences 6-8 in Table 4 are weighted positively (i.e. selected as training instances for the classifier) because they bear some similarity with the expanded reference data. As a result, the test sentences 9-11 are all correctly classified. This examples show the underlying reason why our enhancement of reference data can improve the classification effectiveness.

5 RELATED WORK

5.1 Distantly Supervised Relation Classification

Relation Classification is a fundamental task in natural language processing. Neural network based models have achieved state-of-the-art performance on this task [31, 40, 42, 43]. However, training effective neural classifiers requires a large amount of labeled data, which is usually hard to obtain. Distant supervision provides a way to create massive weakly labeled data for relation classification but the inevitable wrong labels harass reliable training [17, 27].

Most existing studies train relation classifier in DS by applying multi-instance learning (MIL) to reduce the impact of wrong labels [6, 7, 13, 15, 27, 29, 40], which relaxes the relation label of each instance to a *bag* of sentences containing the same entity pair. Assuming at least one sentence within the bag expresses the target relation, MIL-based approaches train and test the relation classification at bag level. They generally face two limitations. (1) They still suffer from noises when all instances within a bag are false [12, 23, 24]. Some recent studies mitigate this issue by incorporating complex attention modules among multi-bags [12, 38]. (2) These MIL based approaches [11, 25] are designed and tested for relation extraction, i.e. to extract all possible relations between a pair of entities, and they are not suitable for sentence-level relation prediction [3, 8]. In our work, we focus on relation classification at sentence level. Nevertheless, we compared with one representative MIL based approach [13] in the experiments.

Alternatively, some recent studies evaluate and select training instances individually without relying on the at-least-one assumption [3, 5, 8, 23, 24, 35, 41], and our work belongs to this family. Previous approaches in this line rely only on the noisy training data to learn instance selection and may suffer from noisy labeling problem. In our work, we introduce a small amount of clean data to guide the instance selection. In addition, previous studies either rely on specific neural networks (e.g. LSTM) [8], or manually crafted reward functions [3, 23, 24], while our approach adopts a model-agnostic meta-learning algorithm without manually specifying any specific form of reward functions or extra models, and thus is more widely applicable in practice.

There are other studies that combine direct supervision and DS. Pershina et al. [22] introduced a small amount of labeled data by leveraging manually selected features from it, while we do not require manually selected features. Beltagy et al. [1] rely on large additional supervised datasets [1] to help identify whether a sentence expresses a relation, while we only use a small set of labeled data as guidance.

A recent work [37] also use manually labeled data for relation classification in DS. They use it to adjust the DS-trained models for relieving the shifted label distribution problem. Our goal is different: we use it to select reliable instances to train effective neural models.

5.2 Gradient-based Meta-learning

Meta-learning [30], also know as learning to learn, has made great progress with recent advances of gradient-based meta-learning [4]. It has wide applications, such as model parameter initialization [4, 18], learning unsupervised update rules [16], learning sample weighting schema [26, 28] and so on.

Some recent studies apply such meta-learning algorithm to learn model parameters for specific relation classification tasks. They focus on supervised relation classiïňĄcation with limited supervision [20] or lifelong relation extraction [19], while we study DS-based relation classification.

Two closely related approaches [26, 28] adopt the same gradient-based meta-learning mechanism as ours to tackle the training set bias issues. They use an ideal validation set which is clean, unbiased and can be scaled to the desired amount to reduce the bias between the training set and test set for image classification tasks. However, this kind of validation set is difficult to obtain in reality. In the DS relation classification task, the available clean reference data is much less than the automatically generated DS training data (the proportion is less than 0.08%), which might lead the model to mode collapse issues. In contrast, by distilling highly confident instances

from noisy data to augment the reference dataset, our approach can leverage more reliable data during training and this effectively improves the performance, as shown in our experiments (Fig 5).

6 CONCLUSION

In this paper, we proposed a meta-learning based approach for distantly supervised relation classification under the guidance of a small set of clean reference data. Our approach learns to reweight instances by minimizing the loss on a dynamically augmented reference set. This process is able to leverage reliable training data and enhances the generalization capability. Experimental results on two distantly supervised datasets show that our approach outperforms previous state-of-the-art noise reduction approaches as well as the meta-reweighting baseline.

Since the reference data in our approach plays as a key role for selecting reliable training instances, in the future, we plan to investigate active learning for annotating some representative reference samples and extend our approach to large scale relation extraction. Other criteria, such as diversity, could also be incorporated in instance selection.

ACKNOWLEDGMENTS

The project is supported by the National Key R&D Program of China (Contract No. 2018YFB21011100) and the National Natural Science Foundation of China (61932001). We would like to thank Jiyang Zhang for his help in the implementation.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Waleed Ammar. 2019. Combining Distant and Direct Supervision for Neural Relation Extraction. In *NAACL-HLT*. 1858–1867.
[2] Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M Strassel, and Jonathan Wright. 2012. Linguistic Resources for 2013 Knowledge Base Population Evaluations. In *TAC*.
[3] Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement Learning for Relation Classification from Noisy Data. In *AAAI'18*.
[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML'17*. JMLR. org, 1126–1135.
[5] Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Denoising Distant Supervision for Relation Extraction via Instance-Level Adversarial Training. *arXiv preprint arXiv:1805.10959* (2018).
[6] 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'11*. 541–550.
[7] 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'17*. 3060–3066.
[8] Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. ARNOR: Attention Regularization based Noise Reduction for Distant Supervision Relation Classification. In *ACL'19*. 1399–1408.
[9] Tingsong Jiang, Jing Liu, Chin-Yew Lin, and Zhifang Sui. 2018. Revisiting Distant Supervision for Relation Extraction. In *LREC'18*.
[10] Pang Wei Koh and Percy Liang. 2017. Understanding Black-Box Predictions via Influence Functions. In *ICML*. JMLR. org, 1885–1894.
[11] Pengshuai Li, Xinsong Zhang, Weijia Jia, and Hai Zhao. 2019. GAN Driven Semi-distant Supervision for Relation Extraction. In *NAACL-HLT*, Vol. 1. 3026–3035. https://doi.org/10.18653/v1/N19-1307
[12] Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2019. Self-Attention Enhanced Selective Gate with Entity-Aware Embedding for Distantly Supervised Relation Extraction. *arXiv preprint arXiv:1911.11899* (2019).
[13] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural Relation Extraction with Selective Attention over Instances. In *ACL'16*, Vol. 1. 2124–2133.
[14] Xiao Ling and Daniel S Weld. 2012. Fine-Grained Entity Recognition. In *AAAI*. 94–100.
[15] Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction. In *EMNLP*. 1790–1795.
[16] Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. 2018. Meta-Learning Update Rules for Unsupervised Representation Learning. *arXiv preprint arXiv:1804.00222* (2018).
[17] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. In *ACL'09*. 1003–1011.
[18] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On First-Order Meta-Learning Algorithms. *arXiv* (2018), arXiv–1803.
[19] Abiola Obamuyide and Andreas Vlachos. 2019. Meta-Learning Improves Lifelong Relation Extraction. *ACL 2019* (2019), 224.
[20] Abiola Obamuyide and Andreas Vlachos. 2019. Model-Agnostic Meta-Learning for Relation Classification with Limited Supervision. In *ACL'19*. 5873–5879.
[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP'14*. 1532–1543.
[22] Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of Labeled Data into Distant Supervision for Relation Extraction. In *ACL'14*. 732–738.
[23] Pengda Qin, Weiran Xu, and William Yang Wang. 2018. DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction. In *ACL'18*, Vol. 1. 496–505.
[24] Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. In *ACL'18*. 2137–2147.
[25] Jianfeng Qu, Wen Hua, Dantong Ouyang, Xiaofang Zhou, and Ximing Li. 2019. A Fine-grained and Noise-aware Method for Neural Relation Extraction. In *CIKM*. ACM, 659–668.
[26] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to Reweight Examples for Robust Deep Learning. In *ICML'18*. 4331–4340.
[27] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *ECML-PKDD'10*. Springer, 148–163.
[28] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting. *arXiv preprint arXiv:1902.07379* (2019).
[29] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *EMNLP'12*. 455–465.
[30] Sebastian Thrun and Lorien Pratt. 1998. Learning to learn: Introduction and overview. In *Learning to learn*. Springer, 3–17.
[31] Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation Classification via Multi-Level Attention CNNs. In *ACL*, Vol. 1. 1298–1307.
[32] Fei Wu and Daniel S Weld. 2007. Autonomously Semantifying Wikipedia. In *CIKM*. ACM, 41–50.
[33] Fei Wu and Daniel S Weld. 2010. Open Information Extraction Using Wikipedia. In *ACL'10*. 118–127.
[34] Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial Training for Relation Extraction. In *EMNLP'17*. 1778–1783.
[35] Kaijia Yang, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Exploiting Noisy Data in Distant Supervision Relation Classification. In *NAACL-HLT*, Vol. 1.
[36] Xuchen Yao and Benjamin Van Durme. 2014. Information Extraction over Structured Data: Question Answering with Freebase. In *ACL'14*, Vol. 1. 956–966.
[37] Qinyuan Ye, Liyuan Liu, Maosen Zhang, and Xiang Ren. 2019. Looking Beyond Label Noise: Shifted Label Distribution Matters in Distantly Supervised Relation Extraction. In *EMNLP-IJCNLP*. 3832–3841.
[38] Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions. In *NAACL-HLT*, Vol. 1. 2810–2819.
[39] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In *EMNLP'15*. 1753–1762.
[40] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation Classification via Convolutional Deep Neural Network. In *COLING'14*. 2335–2344.
[41] Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large Scaled Relation Extraction with Reinforcement Learning. In *AAAI*.
[42] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-Aware Attention and Supervised Data Improve Slot Filling. In *EMNLP'17*. 35–45.
[43] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *ACL'16*, Vol. 2. 207–212.