frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 27 October 2020
doi: 10.3389/fnbot.2020.584192

Boosting Knowledge Base Automatically via Few-Shot Relation Classification

Ning Pang, Zhen Tan, Hao Xu and Weidong Xiao*

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, China

Relation classification (RC) aims at extracting structural information, i.e., triplets of two entities with a relation, from free texts, which is pivotal for automatic knowledge base construction. In this paper, we investigate a fully automatic method to train a RC model which facilitates to boost the knowledge base. Traditional RC models cannot extract new relations unseen during training since they define RC as a multiclass classification problem. The recent development of few-shot learning (FSL) provides a feasible way to accommodate to fresh relation types with a handful of examples. However, it requires a moderately large amount of training data to learn a promising few-shot RC model, which consumes expensive human labor. This issue recalls a kind of weak supervision methods, dubbed distant supervision (DS), which can generate the training data automatically. To this end, we propose to investigate the task of *few-shot relation classification under distant supervision*. As DS naturally brings in mislabeled training instances, to alleviate the negative impact, we incorporate various multiple instance learning methods into the classic prototypical networks, which can achieve sentence-level noise reduction. In experiments, we evaluate our proposed model under the standard *N*-way *K*-shot setting of few-shot learning. The experiment results show that our proposal achieves better performance.

Keywords: knowledge base, relation classification, few-shot learning, distant supervision, multiple instance learning

OPEN ACCESS

Edited by:
Zijun Zhang,
City University of Hong Kong,
Hong Kong

Reviewed by:
Yang Luoxiao,
City University of Hong Kong,
Hong Kong
Lun Hu,
Chinese Academy of Sciences (CAS),
China

****Correspondence:***
Zhen Tan
tanzhen08a@nudt.edu.cn

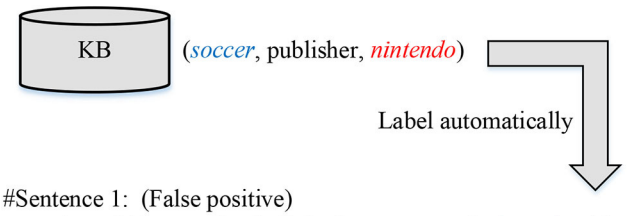
Received: *16 July 2020*
Accepted: *15 September 2020*
Published: *27 October 2020*

Citation:
Pang N, Tan Z, Xu H and Xiao W (2020) Boosting Knowledge Base Automatically via Few-Shot Relation Classification. Front. Neurorobot. 14:584192. doi: 10.3389/fnbot.2020.584192

1. INTRODUCTION

Relation Classification (RC) is defined as identifying semantic relations between entity pairs in given plain texts, which is a crucial task in automatic knowledge base (KB) construction (Bollacker et al., 2008). Mainstream works on this task mainly follow supervised learning, where large-scale and high-quality training data is required (Zeng et al., 2014; Gormley et al., 2015).

However, human-annotated data is always expensive to acquire. Subsequently, recent literature resorted to distant supervision (DS) (Mintz et al., 2009; Riedel et al., 2010) to address the sparsity issue of training data. In DS, it is assumed that *sentences mentioning an entity pair instantiate the relation of the corresponding entity pair in knowledge bases*. With this (untrue) heuristic, large-scale training data can be constructed automatically, but mislabeling is inevitably introduced at the same time. For example, as shown in **Figure 1**, since the triplet (*soccer*, publisher, *nintendo*) exists in a KB, two sentences mentioning the entity pair (*soccer*, *nintendo*) are assigned with the relation "publisher." In fact, the first sentence fails to express the target relation indeed, called false



#Sentence 1: (False positive)
ms. arias said he was fond of playing *soccer* and *nintendo* video games "and anything else that was fun.

#Sentence 2: (True positive)
electronic_arts has been successful with its fifa-branded line of *soccer* games for microsoft, *nintendo* and sony platforms, ⋯

FIGURE 1 | Example of automatic labeling by distant supervision.

Support set:

Relation 1: publisher

#Instance 1: the *diablo* series from *blizzard_entertainment* has done perhaps the best job of combining single-player gameplay with online community.

#Instance 2: *maxis* has gone a step further with *simearth*.

⋯

Relation 2: father

#Instance 1:

⋯

Other Relations

⋯

Query set: (*Soccer*, *nintendo*) Ground truth: publisher

Previous methods:

#Instance: ms. arias said he was fond of playing *soccer* and *nintendo* video games "and anything else that was fun.

Our method:

#Instance 1: ms. arias said he was fond of playing *soccer* and *nintendo* video games "and anything else that was fun.

#Instance 2: electronic_arts has been successful with its fifa-branded line of *soccer* games for microsoft, *nintendo* and sony platforms, ⋯

FIGURE 2 | An example of few-shot relation classification. Previous work selects a single instance as the query, which may fail to provide enough information express the semantic relation between the two entities. To solve the problem, we take multiple instances concerning the same entity pair as the query.

positive instance, while the second one obtains a correct label, which is a true positive instance. Hence, efforts were made to restrain the impact of false positives (Ji et al., 2017; Qin et al., 2018b; Wu et al., 2019). However, these models perform well on common relations, but suffer a dramatic performance drop in classifying long-tail relations, which have few training instances; that is, even though a large amount of training data can be generated by DS, the distributions of such data over different types are unbalanced. Furthermore, they are unable to recognize new relations that have not been seen in training, which potentially restricts their applications in certain scenarios that involve fresh relations in testing.

Lately, pioneering work (Han et al., 2018; Gao et al., 2019) has tried to formulate RC into few-shot learning (FSL) framework (Miller et al., 2000), which aims at accommodating new classes with few examples, while demanding less manual labor than generic supervised learning for fresh relations. Many efforts have made on the few-shot classification task. The early researches fine-tune models which are pre-trained with common classes containing adequate instances by transfer learning (Caruana, 1994; Donahue et al., 2014). After that, metric learning is proposed to project different classes into a distance space (Vinyals et al., 2016; Snell et al., 2017), where similar classes are placed close to each other. Lately, optimization-based meta-learning is developed fast because of its fast-learning ability to learn from previous experience and generalize to new knowledge (Finn et al., 2017; Ravi and Larochelle, 2017). These models, especially prototypical networks, achieve promising results on several benchmarks, but almost all of them focus on image processing. Observing the lack of researches about employing FSL to natural language processing (NLP) tasks, this paper focus on the few-shot relation classification with distant supervision data.

Figure 2 shows an example of few-shot relation classification (FSRC). For an unlabeled query, this method is aimed at classifying it into a correct relation based on a few support instances for each relation. Although FSL requires less training examples in predicting a new relation, moderately large-scale labeled data is necessary to train a promising FSRC model. In specific, a dataset was *manually* labeled for FSRC, namely, FewRel and on top of it, systematic evaluation of state-of-the-art FSL methods (used in computer vision) was carried out for RC (Han et al., 2018). Note that FewRel was constructed by crowdsourcing, and thus, a number much larger than 64×700 of annotations are necessary, where 64 (700, resp.) is the number of relations (labeled instances each relation, resp.) thereof.

1.1. Motivation

To recap, DS can generate large-scale data but suffers from long-tail relations and mislabels; meanwhile, FSRC is able to recognize new relations with few training samples, but requires moderately large amount of human labor for data annotation. Hence, we are at the frontiers of DS and FSRC, being ready to union them, in order to compensate for the downsides of the two paradigms. The combination of DS and FSRC enables the fully automatic method to develop a RC model, which can extract the relation held between two entities. Subsequently, the extracted triplets are used to boost the knowledge base automatically.

In this research, we investigate the task of *few-shot relation classification under distant supervision*. In realization, we refine a previous DS dataset (Zeng et al., 2017), which was built by aligning Wikidata with New York Times corpus, into a reconstructed dataset for FSRC. The DS data is collected automatically, and more details can be seen in Zeng et al. (2017). Taking for granted that the sentences mentioning an entity pair instantiate the relation of the corresponding entities in KBs,

DS data is born with *mislabels* (Riedel et al., 2010). From the example shown in **Figure 2**, we can see that, in the new scenario of distantly supervised FSRC, both support and query sets are practically noisy. If a single false positive instance is sampled as a query like previous studies (Han et al., 2018; Fan et al., 2019; Gao et al., 2019), it cannot be classified into an appropriate relation in the support set. Since a few-shot model is optimized by minimizing the loss of the predictions over the queries, sampling a mislabeled instance as the query will inevitably mislead the optimization process. To tackle this problem, we follow the *at-least-once* assumption, and take an instance bag as a query:

> **Definition 1 (At-least-once assumption)**. *If two entities participate in a relation, at-least one sentence mentioning these two entities express the relation.*
> **Definition 2 (Instance bag)**. *All sentences mentioning a particular entity pair make an instance bag.*

Based on the at-least-once assumption, an instance bag contains enough semantic information to express the relation between the target entities. Therefore, selecting instance bags as queries alleviates the problem of misleading the optimization process which is caused by mislabeled instances. Besides, to alleviate the impact of false positives in the bag, we resort to *multiple instance learning* (MIL) methods, which assigns a single label to the instance bag, and achieve sentence-level noise reduction.

In previous research on FSRC, prototypical networks (PN) achieve promising performance (Han et al., 2018) by measuring the distances between a query and prototypes. The classic approach (Snell et al., 2017) first encodes all instances into a unified vector space, then generates each prototype by averaging all support instances of a relation type. Nevertheless, the mislabeled support instance sampled from DS data may cause a huge deviation for the prototype. In this connection, we conceive a *attention-based MIL* method, which consists of two steps:

- *Denoising the query set*: as discussed above, selecting a single instance that is unfortunately mislabeled as query has negative effects on the optimization of few-shot models. Thus, we take an instance bag as a query which provides enough information for the few-shot models to recognize an appropriate relation concerning an entity pair. Besides, self-attention is supplied to dynamically denoise while producing a more informative query feature vector;
- *Denoising the support set*: for the instances selected as the support set for each relation, to mitigate the issue of substantial deviation of the learned prototype due to mislabeled support instances, support instance-level attention is leveraged to generate a more representative prototype.

In previous studies, Gao et al. (2019) have investigated the support instance-level attention to strengthen the robustness of PN to the noise in support set. However, our work differs from Gao et al. (2019) in two perspectives: (1) Gao et al. (2019) regard the diversity of text as the noise, while in our research, the noise (i.e., mislabeled instances) in the support set is naturally brought in by distant supervision, which is more challenging to be solved; (2) as mentioned above, Gao et al. (2019) select a single instance as the query, which negatively affects the optimization process of few-shot models when distant supervision data is used for training. Differently, we take an instance bag as the query and employ MIL methods to denoise the instance bag. In addition, to evaluate our model on the task of few-shot relation classification under distant supervision, we reconstructed an DS dataset for FSRC.

1.2. Contributions

To sum up, we are among the first to propose to investigate a new task of few-shot relation classification under distant supervision, and the technical contribution is at least three-fold:

- We adapt existing DS data for RC to confront to FSL scenarios, which enables a fully automatic way for FSRC to obtain large-scale potentially-unbiased training data;
- We conceive an attention-based multiple instance learning method over prototypical networks, which reduces noise and emphasizes important signals at both support and query instance levels;
- The proposed task and method are empirically evaluated, and comprehensive results verify the superiority of our proposal over competitors under *N-way K-shot* settings.

1.3. Organization

In section 2, we formally define the task in this work. Related works are discussed in section 3, then we introduce the methodology in section 4. Afterwards, experimental results and detailed analysis are presented in section 5, followed by conclusion.

2. TASK FORMULATION

Formally, the task of *few-shot relation classification under distant supervision* with attention-based MIL is to obtain a function

$$F : (R, S, Q) \rightarrow r, \tag{1}$$

given training data D, which is labeled by existing knowledge bases under the DS assumption. In specific, $R = \{r_1, \ldots, r_i, \ldots, r_m\}$ is the *relation set*, $1 \leq i \leq m$, $m = |R|$, where $|\cdot|$ denotes the cardinality of a set. $D = \{D_{r_1}, \ldots, D_{r_i}, \ldots, D_{r_m}\}$, where D_{r_i} is a set of DS-labeled *instance bags* (all) with relation r_i. S is the *support set*, i.e.,

$$S = \{(s_1^{r_1}, s_2^{r_1}, \ldots, s_{n_1}^{r_1}), \ldots, (s_1^{r_m}, s_2^{r_m}, \ldots, s_{n_m}^{r_m})\}, \tag{2}$$

where relation r_i has n_i support instances, each of which is randomly selected from D_{r_i}. $Q = \{s_1^q, \ldots, s_{|Q|}^q\}$ ($|Q| \geq 1$) is the *query set*, which is essentially an instance bag concerning an entity pair.

The query set Q gives rise to the major difference with respect to the formulation of conventional FSRC in Han et al. (2018), Fan et al. (2019), and Gao et al. (2019). As DS data tends to have

mislabeled instances, if the previous formulation is followed, a mislabeled instance is likely to be sampled as the query instance; in this case, a FSRC model may be intermittently confused during training by the mislabeled query instances, as they substantially deviate from real ones. To overcome the limitation, we propose to employ in training an *instance bag*, instances of which concern the same entity pair and are DS-labeled with the same relation r, to replace a single query instance; by doing this, we expect that the trained model can recover the relation r given its instance bag. Then for testing, the trained model predicts the best relation, where *single or multiple* instances can be supplied.

In this research, we adopt *N-way K-shot* setting, which has been widely used in FSRC (Han et al., 2018; Fan et al., 2019; Gao et al., 2019), i.e., $N = m$, and $K = n_1 = \cdots = n_m$.

3. RELATED WORK

Knowledge base is becoming increasingly important for many downstream applications, and there are various methods to boost the knowledge base (Chen et al., 2019; Zhao et al., 2020). Relation classification (RC) is a vital task for constructing knowledge base automatically. Our work is related to RC via distant supervision (DS) and few-shot learning (FSL). We review the related works as follow.

3.1. Relation Classification Under Distant Supervision

Most existing researches concentrate on neural models via supervised learning (Zeng et al., 2014; Nguyen and Grishman, 2015) or distantly supervised learning (Zeng et al., 2015; Lin et al., 2016). Supervised learning requires a large amount of annotated data, which can be fairly expensive to acquire. As a result, many neural models with supervised learning for RC suffer from data insufficiency (Zeng et al., 2014; dos Santos et al., 2015). DS comes as a remedy (Mintz et al., 2009), which can generate large-scale training data without human labor; whereas, it inevitably brings in mislabels and still has little coverage of long-tail relations. Riedel et al. (2010) formulated RC under DS as a *multiple instance learning* problem to alleviate the influence of mislabels, which achieves remarkable improvement.

On the foundation of this work, other feature-based methods (Hoffmann et al., 2011; Surdeanu et al., 2012) are proposed to better handle noise brought in by distant supervision. Besides, representative neural models include (Zeng et al., 2015; Lin et al., 2016; Feng et al., 2018). Among them, Zeng et al. (2015) perform at-least-one multiple instance learning on DS data. To fully exploit information in an instance bag, Lin et al. (2016) proposed selective attention over instances to dynamically remove noisy samples. Lately, reinforcement learning (Feng et al., 2018; Zeng et al., 2018) and generative adversarial network (Qin et al., 2018a) were combined with these models to further alleviate noise. These models define the task relation extraction as a multiclass classification problem, and they can only extract limited relations as a result. Our work is connected to DS, the major difference is that our proposal is formulated under a FSL framework, which can find new relations in testing, and solve the long-tail relation problem.

3.2. Relation Classification via Few-Shot Learning

Despite satisfactory performance, the aforementioned models show limitations in handling relations with few training samples. FSL provides a feasible solution to the problem of recognizing new classes, which aims at adapting to new classes, given only a few training samples of these classes. Many efforts are devoted to transfer learning methods, which generalizes to new concepts by fine-tuning models pretrained with common classes containing adequate instances (Caruana, 1994; Bengio, 2012; Donahue et al., 2014). Some model-based meta-learning models achieve the rapid learning by designing a special memory unit (Santoro et al., 2016; Munkhdalai and Yu, 2017; Mishra et al., 2018). Another group of studies focus on optimization-based approaches (Finn et al., 2017; Al-Shedivat et al., 2018), which either generate the model parameters directly or predicting the updating gradients for parameters. Afterwards, metric learning is proposed to project instances into a unified feature space, where instances with the same class are placed adjacent with each other (Koch et al., 2015; Vinyals et al., 2016). Prototypical networks used in Han et al. (2018) and Gao et al. (2019), as well as this research is a representative method of metric learning.

As introduced in section 1, Han et al. (2018) first formulated the task of FSRC, and a dataset FewRel for evaluating the task was created via crowdsourcing. Based on FewRel, Fan et al. presented large-margin prototypical networks with fine-grained features (Fan et al., 2019). Wu et al. proposed a dynamic prototypes selection approach with attention to fully capture information in support set (Wu et al., 2020). Seeing texts are more flexible and noisy than images, Gao et al. (2019) devised tailored prototypical networks, distinguishing itself from those used in the area of computer vision. In particular, based on FewRel, noise was introduced by replacing at a certain probability each support instance with a random instance of different relation labels. In contrast, we address the noise issue, i.e., mislabeled instances in specific, which is naturally brought in when investigating the new task of distantly supervised FSRC, which can be comparatively more challenging.

4. METHODOLOGY

As shown in **Figure 3**, the model is established based on prototypical networks (PN), incorporating attention-based multiple instance learning. In this paper, instances are encoded by convolution neural network (CNN) before fed into the PN, and other neural networks can also be employed as encoder.

4.1. Sentence Encoder

This module is used to extract semantic features of an instance. Given a sentence $t = \{w_1, w_2, \ldots, w_n\}$, we first transform the raw text into a low-dimensional embedding representation, and then feed it to neural networks to obtain a feature vector.

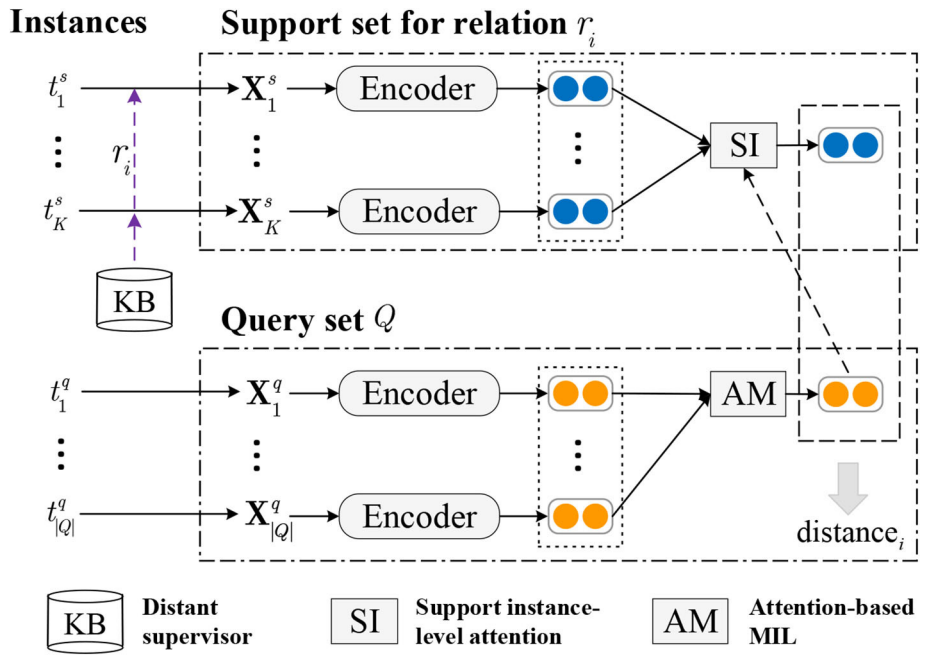


FIGURE 3 | Framework of model for FSRC under DS. t^s and t^q denotes the support instance and query instance respectively, while X^s and X^q are the embedding corresponding representations of t^s and t^q.

4.1.1. Embedding Layer

In our method, we map each discrete word token into a low-dimensional vector by looking up a table of pre-trained word embeddings (Pennington et al., 2014). As thus, a word w_i in the sentence t is converted into a real-valued embedding $\mathbf{w}_i \in \mathbb{R}^{k_w}$, which can express the semantic meaning of w_i. Besides, we also incorporate position features, which have been shown to be useful for RC (Zeng et al., 2014). For each word w_i, it has two relative distances to the two entities. Two position embedding matrices are initialized randomly and each distance can be transformed into a k_p-dimensional vector $\mathbf{p}_i^j \in \mathbb{R}^{k_p}$, $j \in \{1, 2\}$, by looking them up. Then, we concatenate the word embedding and position features as

$$\mathbf{x}_i = [\mathbf{w}_i : \mathbf{p}_i^1 : \mathbf{p}_i^2] \in \mathbb{R}^{k_w + 2k_p}. \tag{3}$$

When gathering the vector representation of all words together, we obtain the input embedding matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. After deriving $\mathbf{X}$, we feed it into a standard CNN for feature extraction.

4.1.2. Encoding Layer

We use convolution neural network (CNN) as the instance encoder, which is of elegant encoding capability and computing efficiency. $\mathbf{X}_{i:j}$ is the concatenation of word vectors $[\mathbf{x}_i : \mathbf{x}_{i+1} : \cdots : \mathbf{x}_j]$. The weight matrix of the sliding filter with a window size of ω is denoted by $\mathbf{W} \in \mathbb{R}^{\omega \times (k_w + 2k_p)}$. The convolution operation is to take a dot production between $\mathbf{W}$ and $\mathbf{X}_{(j-\omega+1):j}$, and generate a vector $\mathbf{c} \in \mathbb{R}^{m-\omega+1}$. Generally, multiple filters are usually required to extract more information, and the corresponding weight matrices are represented by $\widehat{\mathbf{W}} = \{\mathbf{W}_1, \ldots, \mathbf{W}_i, \ldots, \mathbf{W}_d\}$. Each convolution operation can be expressed by

$$c_{ij} = \mathbf{W}_i \otimes \mathbf{X}_{(j-\omega+1):j}, \tag{4}$$

where d is the number of filters, $1 \leq i \leq d$, and $1 \leq j \leq m - \omega + 1$. Afterwards, max-pooling operation is applied on the convolution results to extract the most prominent feature in every dimension, i.e.,

$$y_i = \mathtt{ReLU}(\max_{1 \leq j \leq m} (c_{ij})), \tag{5}$$

where, ReLU is the activation function in our implementation. Hence, a feature vector for an instance $\mathbf{y}^i \in \mathbb{R}^d (i \in \{s, q\})$ is generated by max-pooling layer of CNN, where $\mathbf{y}^s$ (resp. $\mathbf{y}^q$) denotes support (resp. query) instance.

4.2. Attention-Based Multiple Instance Learning Unit

Mislabeled instances are harmful to learning and evaluating queries and prototypes. Hence, we conceive an attention-based multiple instance learning unit to mitigate the impact. **Figure 4** compares the instance selection of our model in both support and query set with the classic method (Snell et al., 2017). From it, we can see that if a false positive instance is sampled as a query, the previous method cannot handle the situation. Besides, the mean selection of support instances is fixed rather than flexible, which restrains the appropriate selection of support instances given a query.

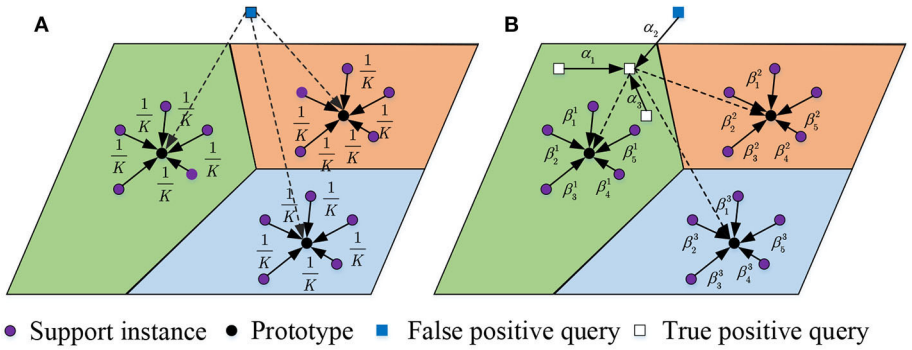


FIGURE 4 | Selection of instances in support and query set. Previous work regards all support instances equally while our method assigns selective attention scores over the support instances. Besides, previous work cannot handle the scenario of false positive queries while our method solves the problem by selecting instance bags as queries. **(A)** Previous method, **(B)** Our method.

4.2.1. Attention-Based MIL Pooling in Query Set

Based on multiple instance learning assumption, in the new formulation, it is of necessity to distinguish the instances of various importance. Consequently, given a set of feature vectors for Q, namely, $\mathbf{Q} = \{\mathbf{y}^q_1, \ldots, \mathbf{y}^q_{|Q|}\} \in \mathbb{R}^{|Q| \times d_c}$, our method leverages an attention-based pooling operation over the multiple instances in the bag. We use a self-attention method (Vaswani et al., 2017), which is defined as

$$\mathbf{E} = \texttt{softmax}(\frac{(\mathbf{Q}\mathbf{W}^1 + \mathbf{b}^1)(\mathbf{Q}\mathbf{W}^2 + \mathbf{b}^2)^\top}{\sqrt{d_c}}), \tag{6}$$

where $\mathbf{W}^1, \mathbf{W}^2 \in \mathbb{R}^{d_c \times d_c}$, and $\mathbf{b}^1, \mathbf{b}^2 \in \mathbb{R}^{d_c}$ are learnable parameters of two linear projection layers, and $\texttt{softmax}(\cdot)$ is the $\texttt{softmax}$ function. $\mathbf{E} \in \mathbb{R}^{|Q| \times |Q|}$ is produced by letting each instance attend mutually. And then, we average each row of $\mathbf{E}$ to generate the attention score for each instance in the query set,

$$\alpha_i = \frac{\exp(e_i)}{\sum_{k=1}^{|Q|} \exp(e_k)}, \tag{7}$$

$$e_k = \frac{\sum_{j=1, j \neq k}^{|Q|} \mathbf{E}_{kj}}{|Q|}, \tag{8}$$

In this way, the selection of query instances is guided by the high-quality ones in the query set.

Then, the query set representation is obtained by consolidating the feature vectors of query instances in a weighted form, i.e.,

$$\hat{\mathbf{y}}^q = \sum_{i=1}^{|Q|} (\alpha_i \mathbf{y}^q_i). \tag{9}$$

In our implementation, we also try other MIL methods, including the maximum pooling over multiple instance, which is defined as,

$$\hat{\mathbf{y}}^q_j = \max_{1 \leq i \leq |Q|} \mathbf{Q}_{ij}. \tag{10}$$

We then concatenate all dimensions and obtain the query feature vector $\hat{\mathbf{y}}^q$.

Another MIL pooling method averages all query instances,

$$\hat{\mathbf{y}}^q = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathbf{y}^q_i. \tag{11}$$

Besides, we also design a perceptron pooling method, which generates the pooling weight for each query instance by,

$$\alpha_i = \frac{\exp(\mathbf{v}^\top \mathbf{y}^q_i)}{\sum_{k=1}^{|Q|} \exp(\mathbf{v}^\top \mathbf{y}^q_k)}, \tag{12}$$

where $\mathbf{v}^\top \in \mathbb{R}^{d_c}$ is a parameter vector. The final query vector can be acquired by Equation (9).

4.2.2. Support Instance-Level Attention

Akin to the case of query set, instances in the support set are not equally useful for learning a prototype when a query set is given. Inspired by Gao et al. (2019), we proceed as follows to get a more informative prototype for each relation,

$$\hat{\mathbf{y}}^s = \sum_{i=1}^{K} (\beta_i \mathbf{y}^s_i), \tag{13}$$

which is a weighted combination of all support instances, and weight β_i is calculated according to the *query set*, as the importance of a support instance varies by queries. Thus, β_i is defined as

$$\beta_i = \frac{\exp(e_i)}{\sum_{k=1}^{K} \exp(e_k)}, \tag{14}$$

$$e_k = \left\| \sigma(\mathbf{y}^s_k \mathbf{W}^s \odot \hat{\mathbf{y}}^q \mathbf{W}^s) \right\|_1, \tag{15}$$

where $\|\cdot\|_1$ is L1-norm, $\sigma(\cdot)$ is a hyperbolic tangent function, $\mathbf{W}^s$ is a learnable parameter matrix, and $\mathbf{y}^q$ is the feature vector of a query generated by Equation (9).

Our proposed attention-based MIL method has two advantages. Firstly, it has *flexibility* in assigning different weights to instances within a query bag, which produces highly informative query vector for bag-level classification. In addition, different attention methods in query set and support set has *interpretability*. High attention weights should be assigned to instances which are true positives, while false positives get low scores.

4.3. Prototypical Networks

The basic idea of prototypical networks is to use prototypes, each generated by a support set, respectively, to delegate a relation. Given a query set Q, distances between its feature vector $\hat{\mathbf{y}}^q$ and all the prototypes are calculated, respectively. Then, the entity pair concerned by the query set is classified as r_i, if the prototype of r_i is of the smallest distance; the probability of Q possessing r_i is

$$p(r_i|Q) = \frac{\exp(-||\hat{\mathbf{y}}_i^s - \hat{\mathbf{y}}^q||_2^2)}{\sum_{j=1}^{m} \exp(-||\hat{\mathbf{y}}_j^s - \hat{\mathbf{y}}^q||_2^2)}. \quad (16)$$

To train the model, we use cross-entropy loss as the target,

$$J(\Theta) = -\sum_j \log p(r_i|Q_j; \Theta), \quad (17)$$

where Θ is the set of parameters used in the model. During model optimization, stochastic gradient decent (SGD) is harnessed to maximize the objective function by updating parameters used in the model iteratively until convergence. For each iteration, mini-batches of samples are selected from the training set.

5. EXPERIMENTS

5.1. Data Preparation and Setup

5.1.1. Dataset

To evaluate the proposed task, we constructed a dataset named DS-Few[1], based on two widely-used DS datasets. The first one was originally built by aligning New York Times corpus with Wikidata[2]. Following the construction method of FewRel, we grouped the instances according to their semantic relations[3], and obtained the *basic version* of DS-Few, consisting of 87 relations with 192,142 instances (61,361 entity pairs) in total[4]. We used 60, 10, and 17 relations for training, validation, and testing, respectively.

For further evaluation, we built an *alternative version* of DS-Few by employing as *test set* another DS dataset—NYT10 (Riedel et al., 2010). Akin to the aforementioned procedure, it was first grouped, and then, we filtered out the sentences that literally appeared in the training set, but retaining the clusters even if the corresponding relations appear in the training set. Eventually, we got a test of 20 relations (10 are seen in the training set) with 142,424 instances in total. In summary, the two versions represent scenarios that are both possible in real life, i.e., a relation may be seen or not during training.

5.1.2. Parameter Setting

For the initial embedding layer of the sentence encoder, we used an embedding set (Wikipedia 2014+Gigaword 5) pretrained by Glove, each of 50 dimensions; for the rest part of the sentence encoder (e.g., position feature and CNN structure), we followed the parameters reported in Zeng et al. (2014). Other parameters were tuned on the validation set. We called stochastic gradient decent to optimize the model, and grid search was harnessed to find the optimal parameters (underscored)—initial learning rate $\lambda \in \{0.01, \underline{0.1}, 0.3, 0.5\}$, and learning rate decay $\gamma \in \{0.01, \underline{0.1}, 0.3, 0.5\}$. That is, λ is multiplied by γ every s training steps. We ran the model on the validation set with $s = 5,000$, every 2,000 iterations for 8 epoches, and the best epoch was chosen for testing. In testing, 3,000 mini-batches are sampled for models to predict, and the prediction accuracy is used as the evaluation metric.

5.1.3. Experiment Setup

We denote our prototypical networks with attention-based MIL as "AMProto." The variants with maximum, average, and perceptron pooling are denoted as "Proto+MAX," "Proto+AVE," and "Proto+PER," respectively. The competitors[5] include SNAIL (Mishra et al., 2018), GNN (Satorras and Estrach, 2018), MAML (Finn et al., 2017) as well as prototypical networks ("Proto"), Proto with self-attention ("Proto+Self") (Wu et al., 2020), and Proto with hybrid attention ("Proto+HATT") (Gao et al., 2019). SNAIL tackles few-shot learning by temporal CNNs with attention mechanism; GNN models each support and query instance as a node in a graph to learn from past experience; MAML optimizes parameters by maximizing the sensitivity of the loss functions of new tasks.

The widely-applied N-way K-shot setting was adopted, $N \in \{5, 10\}$ and $K \in \{5, 10\}$. We tested all models five times, and the *average* results are reported. For fair comparison, we evaluated all models at *instance bag level* (Jiang et al., 2018), i.e., to predict a relation for an instance bag concerning the same entity pair. For the competing approaches that do not work at instance bag level, e.g., Proto (Gao et al., 2019), we trained them at instance level, and chose the instance with the highest confidence score in the instance bag as query in testing.

5.2. Experiment Results

5.2.1. Overall Performance

Table 1 reports the accuracy results of different models. From the results, we would like to highlight that (1) all prototypical networks-based methods exhibit better accuracy than other options (i.e., SNAIL, GNN, and MAML); among the rivals, Proto+HATT is the most competitive since hybrid attention is trained to focus on more important support instances and feature dimensions; (2) AMProto outperforms other prototypical

[1]A download link to the data will be provided in the final version.

[2]https://github.com/thunlp/PathNRE

[3]In this study, clusters with <15 instance bags were discarded, because in the 10-shot settings, at least 15 unique instance bags are necessary—10 as support instance bags and 5 as queries.

[4]For ease of comparison, FewRel is of 100 relations and 70,000 instances.

[5]We omitted comparing with LM-ProtoNet (FGF) (Fan et al., 2019), as (1) all model parameters were unreported, in the original paper, and (2) it performs not as good as Proto+HATT (Gao et al., 2019).

TABLE 1 | Overall results of models.

Methods	Basic version				Alternative version			
	5-way 5-shot	5-way 10-shot	10-way 5-shot	10-way 10-shot	5-way 5-shot	5-way 10-shot	10-way 5-shot	10-way 10-shot
SNAIL	62.46 ± 0.37	68.11 ± 0.28	53.49 ± 0.25	56.22 ± 0.27	61.34 ± 0.35	64.77 ± 0.47	54.35 ± 0.31	57.80 ± 0.36
GNN	63.48 ± 0.59	67.92 ± 0.68	49.07 ± 0.63	54.80 ± 0.61	61.58 ± 0.86	63.28 ± 0.64	50.58 ± 0.74	54.88 ± 0.72
MAML	72.58 ± 0.48	74.46 ± 0.64	56.88 ± 0.41	60.45 ± 0.87	70.37 ± 0.71	73.41 ± 0.56	57.56 ± 0.46	61.97 ± 0.44
Proto	73.03 ± 0.23	75.31 ± 0.18	58.46 ± 0.24	61.86 ± 0.23	71.57 ± 0.32	73.80 ± 0.27	59.55 ± 0.26	62.24 ± 0.21
Proto+Self	73.14 ± 0.31	75.55 ± 0.33	58.51 ± 0.26	62.04 ± 0.25	72.24 ± 0.33	74.63 ± 0.34	59.41 ± 0.28	62.79 ± 0.26
Proto+HATT	73.51 ± 0.11	76.96 ± 0.18	58.85 ± 0.15	63.79 ± 0.17	73.17 ± 0.12	76.60 ± 0.22	59.89 ± 0.12	63.42 ± 0.16
AMProto	74.58 ± 0.21	78.38 ± 0.19	61.51 ± 0.22	65.58 ± 0.18	75.23 ± 0.25	77.86 ± 0.22	62.13 ± 0.17	65.41 ± 0.15

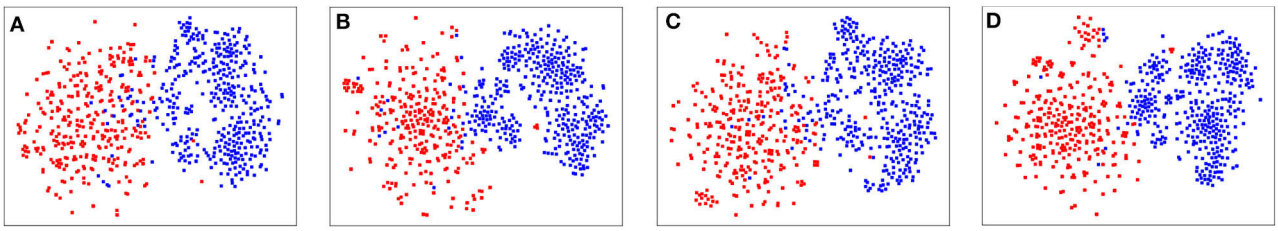


FIGURE 5 | PCA analysis of instance bag embeddings. **(A)** Proto+MAX, **(B)** Proto+AVE, **(C)** Proto+PER, **(D)** AMProto.

TABLE 2 | Comparison of different MIL methods.

Methods	Basic version				Alternative version			
	5-way 5-shot	5-way 10-shot	10-way 5-shot	10-way 10-shot	5-way 5-shot	5-way 10-shot	10-way 5-shot	10-way 10-shot
Proto+MAX	73.48 ± 0.18	76.33 ± 0.14	59.02 ± 0.16	62.24 ± 0.12	73.94 ± 0.17	74.57 ± 0.11	61.19 ± 015	62.82 ± 0.09
Proto+AVE	73.45 ± 0.23	76.52 ± 0.23	59.88 ± 0.19	63.53 ± 0.18	73.89 ± 0.07	75.74 ± 0.12	61.50 ± 0.12	63.25 ± 0.14
Proto+PER	73.17 ± 0.07	77.02 ± 0.15	60.38 ± 0.22	63.63 ± 0.11	73.59 ± 0.13	75.83 ± 0.05	60.66 ± 0.09	63.04 ± 0.05
AMProto	74.58 ± 0.21	78.38 ± 0.19	61.51 ± 0.22	65.58 ± 0.18	75.23 ± 0.25	77.86 ± 0.22	62.13 ± 0.17	65.41 ± 0.15

networks-based FSRC models, implying that it is more robust for the task of FSRC under distant supervision, since it introduces attention-based MIL method to solve the false positive query problem; (3) on two datasets, the corresponding accuracy of different models is quite similar, demonstrating that FSRC models perform similarly on recognizing old relations and fresh relations.

5.2.2. Comparison of Different MIL Methods

To verify the effectiveness of the attention-based MIL, we proceed with comparison analysis by replacing the attention-based MIL with other MIL methods, including MAX, AVE and PER. In this set of experiments, for all models, we keep sampling instance bags as queries in training and testing. The embeddings of these query bags are projected into 2D points by using Principal Component Analysis (PCA), which are shown in **Figure 5**.

The accuracy results are enumerated in **Table 2**. From the results, it reads that (1) the attention-based MIL outperforms all other MIL methods, since self-attention allows the high-quality instances in query bag to guide better instance selection. Besides, due to the interaction between query set and support set, a more informative query feature vector contributes to a more representative prototype. (2) Three competing MIL methods achieve similar performance since they all fail to consider information to guide the assignment of weights over multiple instance in the query bag. The noise contained in query bag also misleads the selection of support instances to form the prototype.

5.2.3. PCA Projection Analysis

This experiment helps appreciate the predictive effect of different MIL methods visually. We conjecture that Proto+MAX, Proto+AVE, and Proto+PER underperform AMProto, due to the selection of high-quality instances in the query bag and the representation of query feature vector. To validate, we randomly selected 400 query instance bags of two arbitrary relations, and encoded them with different models.

(1) there is a subtle difference in the distribution of feature vectors by Proto+MAX, Proto+AVE, and Proto+PER; and (2) in

TABLE 3 | Sample instances in case study.

Relation: parent_taxon	Entity pair in query: (bobcat, lynx)	Attention score
Support instances	① it concludes that the closest living link to the galapagos tortoise, or geochelone nigra, is probably a relatively small tortoise found in South America.	0.16
	② botanically, the poinsettia is euphorbia pulcherrima, a member of the euphorbiacae family, a spurge that comprises about 5,000 specie.	0.27
	③ other examples of convergence include marsupial mammals related to kangaroos and opossums that evolved into creatures resembling lions and wolves.	0.23
	④ a show-stopper was the capra pizza, in which the zing of goat cheese played off beautifully against red and yellow bell pepper slices, black olives and a touch of sage.	0.09
	⑤ by a fluke of nature, a wildcat species—felis silvestris tartessia—has survived unchanged for the past 20,000 years in the mountains of Spain.	0.25
Query instance bag	① bobcats or bobcat tracks have been sighted in the hudson river palisades region, but the lynx rarely ventures south of Northern New England and New York state.	0.18
	② the lynx and the bobcat are similar in size and appearance, although the former's ear tufts are more prominent and its feet larger.	0.28
	③ through a complicated chain of events, it was the bobcat that drove the lynx from New York in the late 1800's.	0.23
	④ … they may have been inspired by the wide footprints of the snowshoe hare and the lynx (a mountain version of the bobcat with especially large feet) in flight.	0.31

The mention in blue and underline is the head entity, while the mention in red and underline is the tail entity.

contrast, feature vectors by AMProto are apt to be linearly-separable when dual attention is exerted; (3) the comparison between AMProto and other MIL methods indicates that the proposed attention-based MIL can learn more distinguishable representation for query bags.

5.2.4. Case Study

We look into the case that AMProto predicts correctly but others fail, to qualitatively show the effectiveness of attention-based MIL. **Table 3** presents a sampled case, where both support instances and query bags are selected from the experiment under 5-way 5-shot setting. Particularly, we presented all instances in support set, and the query instance bag which contains four sentences concerning the entity pair (*bobcat*, *lynx*). The proposed AMProto extracts the relation `parent_taxon` between *bobcat* and *lynx* based on all sentences of the instance bag. In this way, a triplet (*bobcat*, `parent_taxon`, *lynx*) can be formed to complete existing knowledge bases.

It can be seen that our self-attention pooling over the query bag can find the common semantic relation expressed by and distinguish the instances of high attention scores that well express the `parent_taxon` relation, from those of low scores that are mislabeled. Besides, given the query bag, our model can find the high-quality support instances, and assign the lowest attention score to the fourth instance which describes the target relation implicitly.

5.2.5. Results on FewRel Dataset

The two versions of test sets of DS-Few are constructed automatically by distant supervision. To show the performance of few-shot relation classification models on the human labeled data, we also tested our proposed AMProto and all competing methods on FewRel dataset. Specifically, we used the train set of DS-Few to train these models, and the best epochs on the validation set are picked for testing. In our experiments, we tested all few-shot relation classification models on the public train set of FewRel which contains 64 relations. We tested all models on the train set of FewRel due to two reasons: (1) the test set of FewRel is not publicly available; (2) the train set of FewRel contains more relations than the test set (containing 20 relations), which is more challenging for few-shot relation classification models. The results are listed in **Table 4**. From the results, we can read that our proposed AMProto still achieves the best performance among all models when they are tested on the human labeled data.

TABLE 4 | Results on FewRel.

Methods	5-way 5-shot	5-way 10-shot	10-way 5-shot	10-way 10-shot
SNAIL	61.49 ± 0.31	66.43 ± 0.28	54.32 ± 0.36	52.48 ± 0.43
GNN	64.73 ± 0.45	66.62 ± 0.34	50.65 ± 0.37	53.53 ± 0.51
MAML	70.38 ± 0.25	74.52 ± 0.32	60.49 ± 0.24	62.46 ± 0.33
Proto	71.42 ± 0.52	75.44 ± 0.19	61.35 ± 0.28	63.21 ± 0.48
Proto+Self	72.24 ± 0.62	76.39 ± 0.26	62.75 ± 0.22	64.66 ± 0.27
Proto+HATT	72.78 ± 0.24	76.78 ± 0.27	62.47 ± 0.34	65.52 ± 0.26
AMProto	73.85 ± 0.64	77.32 ± 0.36	63.68 ± 0.46	66.27 ± 0.35

5.2.6. Manual Evaluation

When we use the extracted triplets to boost the knowledge base, we usually select those with high confidence scores. It is because we should guarantee the quality of the triplets. Therefore, the precision of top-k triplets (i.e., $P@k$) is an import metric to evaluate few-shot relation classification models. Specifically, we

TABLE 5 | The *P*@*k* values.

Methods	5-way 5-shot				5-way 10-shot			
	P@100	*P*@200	*P*@300	Average	*P*@100	*P*@200	*P*@300	Average
SNAIL	85.00	83.00	79.33	82.44	86.00	84.00	80.67	83.56
GNN	83.00	81.50	79.67	81.39	85.00	85.50	80.33	83.61
MAML	90.00	89.50	86.33	88.61	93.00	91.00	87.67	90.56
Proto	92.00	90.50	87.67	90.06	94.00	91.50	88.67	91.39
Proto+Self	94.00	92.00	88.33	91.44	96.00	93.50	89.67	93.06
Proto+HATT	94.00	92.50	89.33	91.94	97.00	94.00	89.67	93.56
AMProto	95.00	94.00	90.67	93.22	98.00	95.50	91.33	94.94

ranked all extracted triplets according to their confidence scores and calculated the precisions at top-*k* triplets. In our experiments, we tested all models under 5-way 5-shot and 5-way 10-shot settings on the basic version of test set. **Table 5** presents the precisions at top-100, top-200, and top-300. It reads from the results that our proposed AMProto outperforms all baselines. Therefore, it is safer to employ AMProto than other models to boost knowledge base.

6. CONCLUSION

In this paper, to union the advantages of distant supervision and few-shot learning, we have investigated the task of few-shot relation classification under distant supervision. To evaluate, we reconstruct existing distant supervision data to confront the scenario of FSRC. Seeing the unique challenges, we conceive a attention-based multiple instance learning method over prototypical networks to mitigate the mislabeled instances in both support set and query set. Other multiple instance learning approaches, including maximum pooling, average pooling, and perceptron pooling, are selected as our baselines. Empirical study verifies the feasibility of the task and the superiority of the method over other few-shot learning models and various baselines. From the experimental results, we can see that our proposal is more robust to the challenging task.

Our research is evaluated under the classic *N*-way *K*-shot setting of few-shot learning, which can be applied into the scenario of extracting triplets from free texts in designed blanks of forms. However, the real-world application is more complicated. Specially, more free texts may express no relation or other relation not in the support set, which cannot be handled by our proposed model and the competing methods. In the future, we will extend our work to solve the problems of negative instances and cross-domain texts, and enable it to be applicable to more complicated scenario of relation classification.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

NP wrote the paper and conducted the experiments. WX instructed the experiments. HX prepared the data. ZT revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by NSFC under Grant Nos. 61872446, 61902417, and 61701454, NSF of Hunan Province under Grant No. 2019JJ20024, and Postgraduate Scientific Research Innovation Project of Hunan Province (CX20190036).

REFERENCES

Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., and Abbeel, P. (2018). "Continuous adaptation via meta-learning in nonstationary and competitive environments," in *6th International Conference on Learning Representations, ICLR 2018* (Vancouver, BC: OpenReview.net).

Bengio, Y. (2012). "Deep learning of representations for unsupervised and transfer learning," in *Unsupervised and Transfer Learning - Workshop held at ICML 2011* (Bellevue, WA), pages 17-36.

Bollacker, K. D., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008* (Vancouver, BC), 1247–1250. doi: 10.1145/1376616.1376746

Caruana, R. (1994). "Learning many related tasks at the same time with backpropagation," in *Advances in Neural Information Processing Systems 7* (Denver, CO), 657–664.

Chen, Y., Zhao, X., Lin, X., Wang, Y., and Guo, D. (2019). Efficient mining of frequent patterns on uncertain graphs. *IEEE Trans. Knowl. Data Eng.* 31, 287–300. doi: 10.1109/TKDE.2018.2830336

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014* (Beijing), 647–655.

dos Santos, C. N., Xiang, B., and Zhou, B. (2015). "Classifying relations by ranking with convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015* (Beijing), 626–634. doi: 10.3115/v1/P15-1061

Fan, M., Bai, Y., Sun, M., and Li, P. (2019). "Large margin prototypical network for few-shot relation classification with fine-grained features," in *CIKM*, 2353–2356. doi: 10.1145/3357384.3358100

Feng, J., Huang, M., Zhao, L., Yang, Y., and Zhu, X. (2018). "Reinforcement learning for relation classification from noisy data," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)* (New Orleans, LA), 5779–5786.

Finn, C., Abbeel, P., and Levine, S. (2017). "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017* (Sydney, NSW), 1126–1135.

Gao, T., Han, X., Liu, Z., and Sun, M. (2019). "Hybrid attention-based prototypical networks for noisy few-shot relation classification," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI*

Symposium on Educational Advances in Artificial Intelligence, EAAI 2019 (Honolulu, HI), 6407–6414. doi: 10.1609/aaai.v33i01.33016407

Gormley, M. R., Yu, M., and Dredze, M. (2015). "Improved relation extraction with feature-rich compositional embedding models," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015* (Lisbon), 1774–1784. doi: 10.18653/v1/D15-1205

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., et al. (2018). "Fewrel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels), 4803–4809. doi: 10.18653/v1/D18-1514

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. S., and Weld, D. S. (2011). "Knowledge-based weak supervision for information extraction of overlapping relations," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* (Portland, OR), 541–550.

Ji, G., Liu, K., He, S., and Zhao, J. (2017). "Distant supervision for relation extraction with sentence-level attention and entity descriptions," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, CA), 3060–3066.

Jiang, T., Liu, J., Lin, C., and Sui, Z. (2018). "Revisiting distant supervision for relation extraction," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018* (Miyazaki).

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop, Vol. 2* (Lille).

Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016). "Neural relation extraction with selective attention over instances," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016* (Berlin). doi: 10.18653/v1/P16-1200

Miller, E. G., Matsakis, N. E., and Viola, P. A. (2000). "Learning from one example through shared densities on transforms," in *2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000)* (Hilton Head, SC), 1464–1471. doi: 10.1109/CVPR.2000.855856

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). "Distant supervision for relation extraction without labeled data," in *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Singapore), 1003–1011. doi: 10.3115/1690219.1690287

Mishra, N., Rohaninejad, M., Chen, X., and Abbeel, P. (2018). "A simple neural attentive meta-learner," in *6th International Conference on Learning Representations, ICLR 2018* (Vancouver, BC).

Munkhdalai, T., and Yu, H. (2017). "Meta networks," in *Proceedings of the 34th International Conference on Machine Learning, Vol. 70* (Sydney, NSW: JMLR. Org), 2554–2563.

Nguyen, T. H., and Grishman, R. (2015). "Relation extraction: perspective from convolutional neural networks," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. (Denver, CO) 39–48. doi: 10.3115/v1/W15-1506

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014* (Doha), 1532–1543. doi: 10.3115/v1/D14-1162

Qin, P., Xu, W., and Wang, W. Y. (2018a). "DSGAN: generative adversarial training for distant supervision relation extraction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018* (Melbourne, VIC), 496–505. doi: 10.18653/v1/P18-1046

Qin, P., Xu, W., and Wang, W. Y. (2018b). "Robust distant supervision relation extraction via deep reinforcement learning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018* (Melbourne, VIC), 2137–2147. doi: 10.18653/v1/P18-1199

Ravi, S., and Larochelle, H. (2017). "Optimization as a model for few-shot learning," in *5th International Conference on Learning Representations, ICLR 2017* (Toulon).

Riedel, S., Yao, L., and McCallum, A. (2010). "Modeling relations and their mentions without labeled text," in *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010* (Barcelona), 148–163. doi: 10.1007/978-3-642-15939-8_10

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. P. (2016). "Meta-learning with memory-augmented neural networks," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016* (New York City, NY), 1842–1850.

Satorras, V. G., and Estrach, J. B. (2018). "Few-shot learning with graph neural networks," in *6th International Conference on Learning Representations, ICLR 2018* (Vancouver, BC).

Snell, J., Swersky, K., and Zemel, R. S. (2017). "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017* (Long Beach, CA), 4077–4087.

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (2012). "Multi-instance multi-label learning for relation extraction," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012* (Jeju Island), 455–465.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017* (Long Beach, CA), 5998–6008.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016* (Barcelona), 3630–3638.

Wu, L., Zhang, H.-P., Yang, Y., Liu, X., and Gao, K. (2020). "Dynamic prototype selection by fusing attention mechanism for few-shot relation classification," in *Asian Conference on Intelligent Information and Database Systems* (Springer), 431–441. doi: 10.1007/978-3-030-41964-6_37

Wu, S., Fan, K., and Zhang, Q. (2019). "Improving distantly supervised relation extraction with neural noise converter and conditional optimal selector," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (Honolulu, HI), 7273–7280. doi: 10.1609/aaai.v33i01.33017273

Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015* (Lisbon), 1753–1762. doi: 10.18653/v1/D15-1203

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). "Relation classification via convolutional deep neural network," in *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers* (Dublin), 2335–2344.

Zeng, W., Lin, Y., Liu, Z., and Sun, M. (2017). "Incorporating relation paths in neural relation extraction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017* (Copenhagen), 1768–1777. doi: 10.18653/v1/D17-1186

Zeng, X., He, S., Liu, K., and Zhao, J. (2018). "Large scaled relation extraction with reinforcement learning," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, (*AAAI-18*), *the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)* (New Orleans, LA), 5658–5665.

Zhao, X., Zeng, W., Tang, J., Wang, W., and Suchanek, F. (2020). An experimental study of state-of-the-art entity alignment approaches. *IEEE Trans. Knowl. Data Eng*. doi: 10.1109/TKDE.2020.3018741

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.