ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.003

A Framework for Domain-specific Distant Supervised Named Entity Recognition

Long Qin [1], Xiaoge Li [1]

[1] Xi'an University of Posts and Telecommunications

Abstract. With the development of knowledge graphs in the industrial field, Constructing KGs from domain-specific problems is greatly important. However, there is no consensus in regard to a plausible and definition of entities and relationships in the domain-specific knowledge graph. Further, in conjunction with several limitations and deficiencies, various domain-specific entities and relationships recognition approaches are far from perfect. Specifically, named entity recognition in domain-specific is a critical task for the natural language process applications. However, a bottleneck problem with named entity recognition in domain-specific is the lack of annotated data. To address this challenge, A domain-specific distant supervised named entity recognition framework is proposed. The framework is divided into two stages: first, the distant supervised corpus is generated based on the entity linking model of graph attention neural network; secondly, the generated corpus is trained as the input of the distant supervised named entity recognition model to train to obtain named entities. The link model is verified in the ccks2019 entity link corpus, and the F1 value is 2% higher than that of the benchmark method. The re-pre-trained BERT language model is added to the benchmark method, and the results show that it is more suitable for distant supervised named entity recognition tasks. Finally, it is applied in the computer field, and the results show that this framework can obtain domain named entities.

Keywords: distant named entity recognition, entity linking, knowledge graph, graph attention neural network

1. Introduction

The Knowledge Graphs (KGs) depicts an integrated collection of real-world entities which are connected by semantically-interrelated relation. Specifically, KGs plays a greatly important role in industry, such as agriculture, health care, stock and so on. Therefore, KGs always to be used as the main means of tacking a plethora of real-life problems in various domains. However, there is no consensus in regard to a plausible and inclusive of a domain-specific entities and relationships. Named entity recognition (NER) and relation extraction (RE) are two basic tasks in natural language processing and attracts more and more attention. NER as the basis of RE is very important. Domain NER can be described as the linguistic representations of domain specific concepts. NER is the task of detecting mentions of real-world entities from text and classifying them into predefined types. Although NER have achieved great process in the task of natural language processing, it still faces significant obstacles regarding domain named entity recognition. In recent years, more and more researches focus on NER in deep learning methods. With the development of deep learning, deep learning models have shown strong performance. However, most deep learning methods rely on large amounts of labeled training data. To tackle the label scarcity issue, the distant supervised named entity recognition is proposed. The labeling procedure is to match the tokens in the target corpus with concepts in knowledge bases in distant supervised named entity recognition. Nevertheless, distant supervised named entity recognition suffered from two major challenges: incomplete annotation and label noise issue. To solve above problems of distant supervision for NER, some researches have attempted to address it. Yang [1] adopt the partial annotation CRFs to consider all possible labels for unlabeled tokens, but this methos still require a considerable amount of annotated tokens or external tools; Cao [2] attempt to induce labels for entity mentions based on their occurrence popularity in the concept taxonomy, which can suffer from labeling bias and produce mislabeled data; Liang [3] studied the open-domain named entity recognition problem under distant supervision, which uses the pre-trained language model to improve the prediction performance of the named entity recognition model. With the large-scale pre-training language model rapidly becoming the mainstream method of natural language processing tasks [4]. The attention mechanism

proposed by Vaswani [5] establishes the position of BERT pre-training language model in natural language processing tasks [6].

Domain-specific named entity recognition, also often referred to as Automatic term extraction (ATE), is the automated process of identifying terms in specialized texts. In recent years, it has become an important pre-processing step in many natural language processing tasks. Wang [7] used two deep learning classifiers to extract terms; Amjadian [8] proposed an efficient method of combining distributed representation with term extraction; Hatty [9] constructed classification tasks by defining fine-grained terms and using basic neural networks; Kucza [10] used recurrent neural network to term extraction by means of sequence labeling; Shah [11] used unsupervised learning to extract terms in material science; Sajatovic [12] proposed a topic modeling method to extract terms on individual documents; Kessler [13] extracted architectural terms based on search engines and Wikipedia; Pollak [14] proposed a method of extracting terms from literary corpus and automatically aligned terms to from a domain term dictionary; Terry [15] proposed a new method of monolingual and multilingual labeling to generate dataset for downstream tasks; Terry [16] once again introduced the search platform in the field of automatic term extraction, and made a detailed introduction to automatic term extraction; Kafando [17] proposed an intelligent method for extracting biomedical terms from text documents and subsequent analysis, which combines statistical metrics and syntactic change rules to extract term variants from the corpus.

Noisy annotation is one of the major challenges in domain-specific distant supervised named entity recognition. Entity alignment (EA) aims to reduce the problem of noisy annotation in NER. The same entity mention can be mapped to multiple entity types in the knowledge bases. For instance, the entity mention ‘Machine learning’ can be mapped to both ‘a book’ and ‘subject’ in the knowledge base. While existing methods can lead to many false-positive samples and hurt the performance of named entity recognition models. To solve noisy annotation in domain-specific distant named entity recognition and a large scale of labeled corpus, we propose our model, a domain named entity recognition with distant supervision, which combine with the task of automatic term extraction and distant supervised named entity recognition. The method obtains labeled corpus by linking Wikipedia and domain paper entities, and applies the corpus to the distant supervised named entity recognition model.

In summary, we make the following contributions:

1)We propose a framework of domain-specific distant supervised named entity recognition, and this framework is applied to obtain named entities in computer field.

2) We use graph attention neural network to solve the problem of entity linking errors caused by multiple mappings of the same entity in the task of distantly supervised named entity recognition.

3)We demonstrate that re-pre-trained language model can also provide additional semantic information during the training process for distantly supervised named entity recognition.

2. Approach

Our approach is used to extract domain-specific named entitles (see Figure 1). Specifically, we propose a two-stage training algorithm: In the first stage, Generation of distant labeled data based on graph attention neural network. In this stage, firstly, the entities of Wikipedia are extracted, and the entity information is used as the knowledge base. Then the entity graph is constructed based on the knowledge base. Secondly, the BERT language model is used as the input to extract the graph embedded representation of the entity nodes in the graph attention neural network, and then the text information and entity graph embedded representation information of the domain paper are used as features to carry out the entity linking task in the binary entity link model based on BERT. In the second stage, the results of entity linking are used as a small amount of labeled corpus, and the results of entity links are trained as entity dictionaries to generate labeled data by matching with a large number of unlabeled data as the input of distant supervised named entity recognition. We give more details on our approach in what follows.

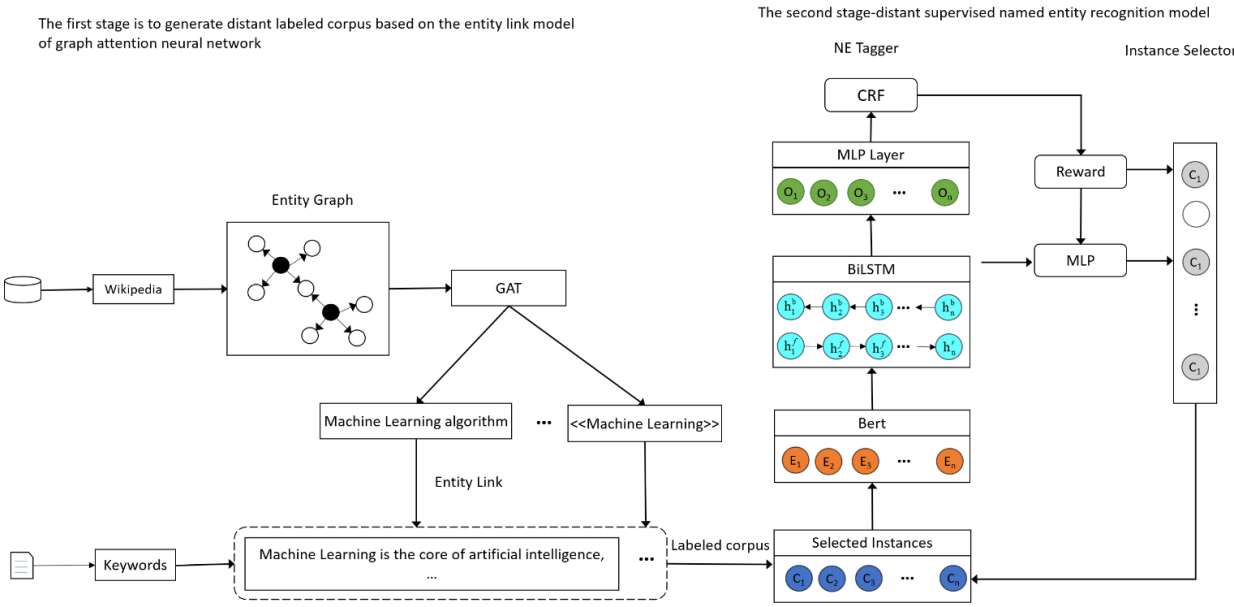

Fig. 1. The framework of domain distant supervised named entity recognition. We propose a two- stage training algorithm: In first stage, combine Wikipedia and domain paper to generate labeled data based on the entity link model of graph attention neural network; In the second stage, a distant supervised named entity recognition.

2.1. Entity Linking Model Based on Graph Attention Neural Network

(1) Entity extraction based on Wikipedia classification system

Wikipedia uses different classification methods for each classification object. Such as for the contents of general items are classified according to disciplines, for chronological accounts are classified by time, for character items use nationality and occupation. From the perspective of discipline knowledge classification system. The entries of Chinese Wikipedia are mainly divided into eight categories: Religion and belief, around the world, Humanities and Social Sciences, Life art and culture, Engineering technology and Applied Science, Nature and Natural Science, Chinese culture and Sociology. Each directory is subdivided into many secondary directories, and so on, until it is subdivided into specific items. Entries in the computer field can be obtained from the secondary directory computer science under engineering technology and applied science. Starting from computer science, the sub categories and subpages under this classification are recursively extracted to obtain the entity dictionary in the computer field as the named entity.

(2) Entity Graph

The entities in the computer field are extracted in Wikipedia, and the basic information of the entities is obtained from Baidu Encyclopedia, then the entity knowledge base is constructed by using the entity knowledge base.

Table 1. Relationship Definition of Entity

Relationship type	triple	example
alias	<entity, alias, value >	<GAN, alias, Generative Adversarial networks>
attribute	<entity, attribute, value>	<GAN, domain, deep learning >

Table 2. A Piece of Knowledge Base Information

{"alias": "[Gan]", "subject id": "10000", "subject": "Gan", "type": "computer", "data": [{"predicate": "Chinese name", "object": "generative countermeasure network"}, {"predicate": "foreign name", "object": "Generative Adversarial Networks"}, {"predicate": "abbreviation", "object": "GAN"}. {"predicate": "domain", "object": "deep learning"}, {"predicate": "composition", "object": "generating model and discriminant model"} {"predicate": "summary", "object": "generative countermeasure network (GAN, Generative Adversarial Networks) is a deep learning model. It is one of the most promising methods of unsupervised learning in complex distribution in recent years. The model produces a good output through the mutual game learning of (at least) two modules in the framework: the generative model) and the discriminative model. In the original GAN theory, G and D are not

required to be neural networks, but only functions that can fit the corresponding generation and discrimination. But in practice, depth neural network is generally used as G and D. " }, {"predicate": "keyword", "object": "Neural network"}, {"predicate": "keyword", "object": "Unsupervised learning"}}

The information of the entity is represented by the graph structure, and the graph embedded representation between the entity and its attributes is obtained through the graph attention network. This computer domain entity knowledge base is constructed by referring to the CCKS2019 entity link knowledge base, which include entity and its attribute value, subject-entity, subject-id, subject-type, alias, predicate and object value. The entity triples are defined as follows Table1 and a piece of knowledge base information as Table 2.

The candidate entity node is taken as the central node, and the alias and attribute values are used as neighbor nodes to describe the knowledge of the central node. The entity graph of computer domain is constructed by using the defined triple. At the same time, in order to reduce the segmentation of the subgraph, all the description attributes of the candidate entities of each data in the knowledge base are spliced with the attribute value text, and the keywords are extracted by TF-IDF. If the extracted keywords coincide with the original attributes, only one is retained. The keywords are integrated into the entity graph (see Figure 2 and Table 3).

Table 3. Triple of Computer Entity

entity	attribute	attribute value
GAN	subject id	10000
GAN	Chinese name	生成对抗式网络(generative adversarial networks)
GAN	foreign name	generative adversarial networks
GAN	abbreviation	GAN
GAN	domain	deep learning
GAN	composition	generating model and discriminant model
GAN	summary	text
GAN	keyword	neural network
GAN	keyword	unsupervised learning

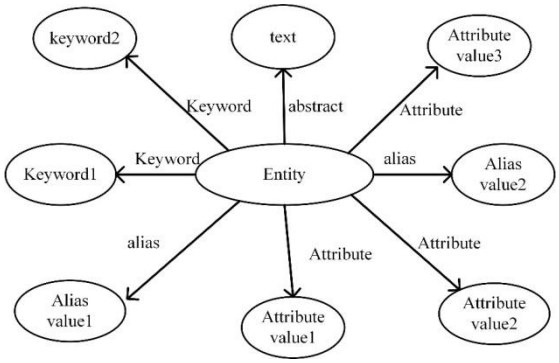

Fig. 2. Computer entity graph

(3) Graph attention neural network

The data of graph structure contains two features: vertex feature and neighbor feature. GCN [18] cannot handle the problem of dynamic graphs, and it is not easy to assign different weight to different neighbors. However, GAT consider the structure of the graph in space, so it can perfectly adapt to the dynamic graph. Velickovic [19] proposed graph attention neural network, and many researchers have applied it to a variety of natural language processing tasks and achieved good results; Zhang [20] build graph information for documents and obtain fine-grained word representations of their local structures for text classification; Huang [21] used graph neural network for text classification; Zheng [22] proposed a new multi-granularity

machine reading comprehension framework in machine reading comprehension, which uses graph attention network to obtain different levels of representation so that they can be learned at the same time.

The entity graph is constructed according to the triple relationship of entities and their attributes defined in Table 1 and the keywords added in figure 2. The entity graph is studied by embedding the graph into the node representation through the graph attention neural network. Using the attention mechanism, the characteristics of neighboring nodes are weighted and summed, and different weights are assigned to different neighboring nodes according to their characteristics. The graph attention neural network model is as Figure 3.

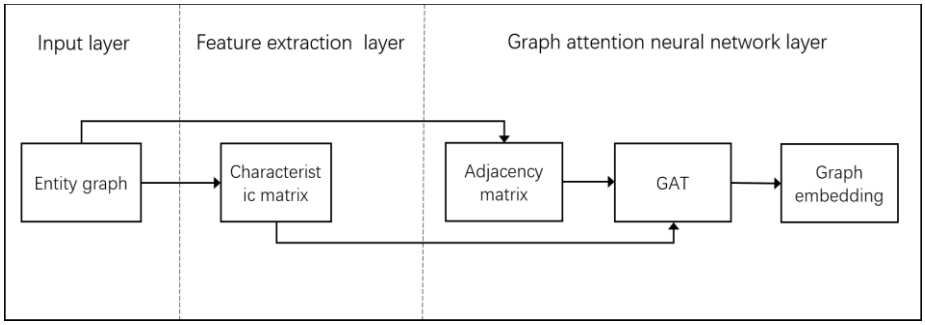

Fig. 3. Graph attention mechanism Neural Network and its graph embedding.

The model consists of input layer, feature extraction layer and graph attention neural network layer.

The input layer is entity graph, which consists of a set of X nodes $V=\{V_1,V_2...V_X\}$ and Y edges $E=\{E_1,E_2...E_Y\}$. In the feature extraction layer, the BERT pre-training model is used to extract the feature of the node text in the entity graph, and the $\mathrm{m}\times\mathrm{n}$ feature matrix is obtained, where m is the number of graph nodes and n is the feature dimension. The graph attention layer obtains the feature matrix and the adjacency matrix from the input layer and the feature extraction layer, and inputs them to the two-layer graph attention neural network for representation. The final output is the graph node feature representation represented by GAT.

The following is a detailed description of the GAT model.

The input to GAT is a set of node features, $h=\{\vec{h}_1,\vec{h}_2...\vec{h}_N\}$, $h_i\in\mathbb{R}^F$, where N is the number of nodes, and F is the number of features in each node. $h'=\{\vec{h}'_1,\vec{h}'_2...\vec{h}'_N\}$, $\vec{h}'_i\in\mathbb{R}^F$, as its output.

The coefficients computed by the attention mechanism may be expressed as:

$$\alpha_{ij}=softmax\left(\sigma\left(\vec{a}^T\left[W\vec{h}_i\,\|\,W\vec{h}_j\right]\right)\right) \tag{1}$$

Where α_{ij} represents the attention coefficient between nodes i and j, $W\in\mathbb{R}^{F\times F'}$, is a weight matrix, $\vec{h}_i$ and $\vec{h}_j$ represents the node characteristics of node i and j, the dimension of $\vec{h}_i$ is $1\times F$, then the dimension of $w\vec{h}_i$ is $1\times F'$, $\|$ represents concatenation, and the tensors of two $1\times F'$ are glued together to a large tensor of $1\times 2F'$. Then by multiplying the dimension of the attention convolution kernel coefficient $\vec{\alpha}^T$, and the dimension of $\vec{\alpha}^T$ is $2F'\times 1$,the final result is a number. σ indicates the activation function, which is LeakyReLU. After activating the function, the final attention result is calculated by softmax.

$\vec{h}'_i$ comes from the following formula:

$$\vec{h}'_i=\sigma\left(\sum_{j\in\mathrm{N}_i}\alpha_{ij}W\vec{h}_j\right) \tag{2}$$

$\vec{h}_i$ represents the output characteristics of the layer with respect to node i, where N_i is some neighborhood of node i in the graph. Specifically, multi-head attention to be beneficial for mechanism. Where k independent attention mechanisms execute the transformation of Equation 2, and then their features are concatenated, so the output feature representation:

$$\vec{h}'_i=\Big\|_{k=1}^{\mathrm{K}}\sigma\left(\sum_{j\in\mathrm{N}_i}\alpha_{ij}W\vec{h}_j\right) \tag{3}$$

The final output layer can be represented as:

$$\vec{h}'_i=\sigma\left(\frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\sum_{j\in\mathrm{N}_i}\alpha_{ij}{}^kW^k\vec{h}_j\right) \tag{4}$$

(4) Entity link

A large number of researchers have improved the BERT model in the task of entity linking. Zhan [23] proposed to introduce the BERT pre-training language model into the entity link task to analyze the context

of the entity reference item and the relevant information of the candidate entity. By improving the effect of semantic analysis to enhance the results of entity links, and using TextRank keyword extraction technology to enhance the topic information of target entity comprehensive description information, and enhance the accuracy of text similarity measurement; Luong [24] combine global and local attention mechanism to obtain the hidden state of the text.

In 2019, the National Conference on knowledge Graph and semantics issued the task of entity linking evaluation for Chinese short texts. The best solution uses dictionary matching to get the entities in the short text, and finally uses the BERT-EntityNameEmbedding (BERT-ENE)[1] model to filter the results, so as to achieve entity recognition. In the part of entity link, the two-classification model based on BERT is used to predict and rank the candidate entities.

The entity linking model of our framework is shown in figure 4, using binary entity links based on BERT. We consider node feature representation of entities in graph attention mechanism network. Short text and description text of the entity to be lined as input of model. The features are the BERT vector of the text, the start and end position vectors of the candidate entity, and the node feature representation of the entity in the graph attention mechanism neural network. The four feature vectors are spliced, and after the full connection layer, the probability scores of the candidate entities are sorted are sorted by sigmoid activation, and the correct entity with the highest probability is selected.

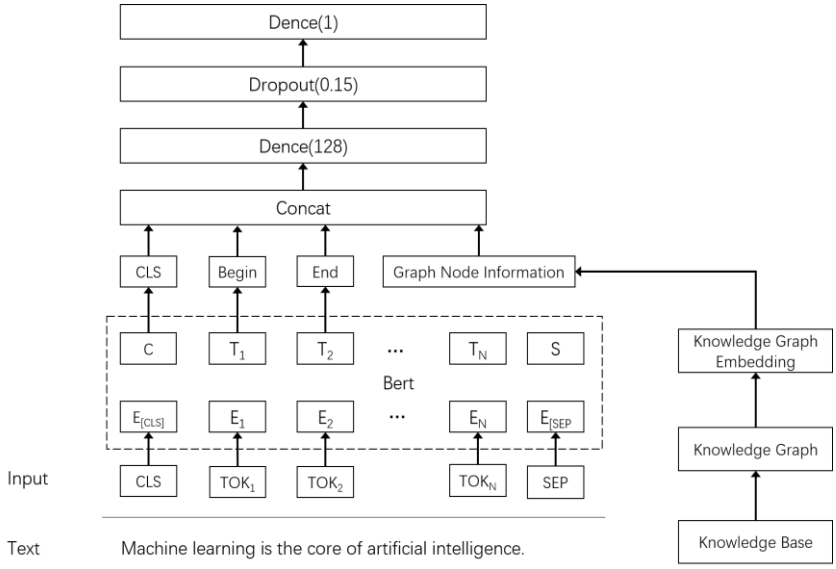

Fig. 4. Entity link model based on BERT and Graph embedding.

2.2. Named Entity Recognition with Distant Supervised

The Chinese language pre-training model has strong performance in natural language processing tasks. Gururangan [25] proposed that language model pre-training can greatly improve the effect of subtasks. For example, continuing pre-training on the data set of specific tasks can improve the effect very cheaply. The effect can be improved by continuing pre-training on the data set of the target domain, and the more irrelevant the corpus of the target domain and the original training corpus is, the more obvious the improvement effect is.

A number of pre-training models have also been produced in the field of Chinese pre-training models, such as the Chinese BERT model[2], the Roberta pre-training model released by IFLYTEK of Harbin University of Technology[3], and the ERNIE pre-training model released by Baidu[4].

In order to study the performance of the pre-training model in the task of distantly supervising named entity recognition, on the basis of three Chinese pre-training models, People's Daily corpus, Amazon commodity review corpus and restaurant review corpus are used for retraining.

[1] https://github.com/panchunguang/sskc_baidu_entity_link
[2] https://huggingface.co/bert-base-chinese/tree/main
[3] https://github.com/ymcui/Chinese-BERT-wwm
[4] https://github.com/nghuyong/ERNIE-Pytorch

After obtaining the computer domain entity, the labeled corpus is obtained by matching the entity and unlabeled corpus, which is applied to the distant supervision of named entity recognition task (see figure 1).

Following Yang et al. (2018)[5], we consider BERT pre-training language model on the distant supervised entity recognition model. The following is a detailed introduction of this model.

As shown in Figure 1, the model of distant supervised named entity recognition consists of two modules: the NE Tagger built on the idea of partial annotation learning to reduce the effect of unknown-type characters, the instance selector which choose positive sentences from a large scale of unlabeled corpus and provides them to the NE tagger to train model.

Initially, we get a small set of labeled data D from the entity linking model and a large scale of unlabeled data U. According to the results of entity linking, we collect named entity to construct dictionary E about computer field, then using the entries of E to match the sentences in U by the method of distant supervision. And we also obtain a set of sentences containing at least matched string, and the set is called I. According to the traditional BIO schema to represent the tags of sentences, the beginning character of an entity in I are marked with "B-XX", "I-XX" is used to mark other characters of the entity, and the character as "O" if it is not in the entity.

LSTM-CRF-PA

It is a common problem known as false negative instance in distant supervised named entity recognition. If we arbitrarily label as "O" which may misguide model to learn the false instance. Therefore, each non-matched character should be tagged as the appropriate label. A set of label sequences z for every distantly supervised sentence, whose probability is naturally the sum of probability of each possible label sequence $\tilde{y}$ in z. Therefore, the probability of the distantly supervised instance is calculated as:

$$p(z|x) = \sum_{\tilde{y} \in z} p(\tilde{y}|x) = \frac{\sum_{\tilde{y} \in z} e^{score(x, \tilde{y})}}{\sum_{\tilde{y} \in Yx}^{n} e^{score(x, \tilde{y})}} \tag{5}$$

The loss function of the model with CRF-PA cam be computed as follows:

$$loss(\theta, x, z) = -\log p(z \mid x) \tag{6}$$

Instance Selector for Noisy Annotation

To train an agent as an instance selector with reinforcement learning technology. The initial labeled data D and the distantly supervised data I is denoted as a candidate dataset A. At each episode, we collect a random-size package of instances B from A. By default, all the supervised instances in the current package are selected without decisions of agent. For each distantly supervised instances in the current package, the agent performs an action from the {1,0} to decide whether to select this instance. The agent will be rewarded when all actions are completed. The reward represents action feedback on this package and will be used to update the agent.

State representation

The vector S_t represents the current instance and its label sequence, which consists of two information: first, the vector representation of the output from BiLSTM layer. Secondly, the label score calculated with output of the MLP layer from the shared encoder and annotation of this instance.

Policy network

The agent determines whether the behavior selector will select the t-th distantly supervised instance. Then we use a logistic function as the policy function:

$$A_\theta(s_t, a_t) = a_t \sigma(W * S_t + b) + (1 - a_t)(1 - \sigma(W * S_t + b)) \tag{7}$$

Reward

The reward is used to evaluate the ability of current NE tagger to predict labels of each character. The model receives a delayed average reward when it completes all elections in current package, and before that

[5] https://github.com/rainarch/DSNER

the reward for each action is zero. The current package A consists of two subsets: The labeled data D from entity linking and B from the distantly supervised instances. The NE tagger calculates the probability of each sentence in A. The reward can be calculated on selected distantly supervised instances $\tilde{A}$ and the labeled data:

$$r = \frac{1}{|\tilde{A}_s| + |\tilde{H}|}\left(\sum_{x_j, z \in \tilde{A}_s} \log p(z \mid x_j) + \sum_{x_j, z \in \tilde{H}} \log p(y \mid x_k)\right) \tag{8}$$

Selector training

In order to maximize the reward of the selections, we use policy gradient method to optimize the policy network. For each random-size package A, the feedback for each action is the same as the average reward r. Then we calculate the gradient and update the selector.

$$\theta = \theta + \alpha \sum_{t=1}^{|\tilde{A}|} r(a_t) \nabla_\theta \log A_\theta(s_t, a_t) \tag{9}$$

3. Experimental Results

3.1. Datasets

Entity linking. We use the evaluation data set provided by the CCKS2019 entity linking task for Chinese short text as the verification data of the entity linking model. Its dataset includes 90000 pieces of labeled data and 39925 pieces of knowledge base information.

Distant supervised named entity. Following Yang [1] , the news corpus dataset is selected, 3000 sentences were randomly selected as the training corpus, 3328 as the verification corpus and 3186 as the test corpus. In order to verify the effect of corpus retraining on BERT pre-training model, People's Daily 2016 data set, Amazon product review data and food review data were used as pre-training corpus and retrained on three pre-training models: BERT, Roberta and ERNIE.

Computer domain entity linking and distant supervised named entity recognition. Wikipedia is used to build a computer domain entity knowledge base, which contains a total of 84493 pieces of data. The entity link data is the abstract of the paper, and 201608 pieces of data are extracted.

After linking through the entity, the named entity is matched in the untagged corpus to get 9000 pieces of data, which are divided into training, verification and testing distant supervised named entity recognition model according to 1:1:1.

3.2. Entity Linkong Experiments

(1) As shown in

In the entity linking experiment, the entity linking task of ccks2019 is used to verify. On the basis of the first prize in the ccks2019 contest, the BERT-ENE model of the original author is adopted in the entity recognition module. In the entity linking module, the knowledge base information is transformed into graph information, and the link model is realized by using short text and entity embedding based on graph attention mechanism. The results are shown in the following table2.

Table 4. Result of Entity Link (%)

model	F1
BERT	80.13
BERT-ENE	83.29

As can be seen from Table 4, the effect of adding entity graph embedding model is better than that of text input based on BERT. It shows that through graph attention neural network, we can obtain more semantic information about the entity, its attributes and its text.

3.3. Distant Named Entity Recognition Experiments

A distant supervised named entity recognition model based on BERT pre-training language model is adopted. In order to compare with the word2vec of the benchmark model, the first 100-dimensional vector of the output of the first layer of BERT is selected as the feature. And under the People's Daily corpus, Amazon commodity review corpus and food review corpus, the Chinese BERT model, Roberta model and ERNIE model are pre-trained again. The retraining corpus training set uses data about the size of 300MB.

Table 5. Result of Word2vec and Chinese BERT

model	DEV			TEST		
	P	R	F1	P	R	F1
wordvec-10	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	**89.64**	**80.73**	**84.95**	**85.48**	**77.39**	**81.23**

As shown in Table 5, the performance of the BERT pre-training language model is obviously better than that of the word2vec model. The effect of retraining the three Chinese pre-training language models on the People's Daily corpus is shown in Table 6.

TABLE 6. Results of Re-pre-training Model on the People's Daily Corpus

model	DEV			TEST		
	P	R	F1	P	R	F1
word2vec-100	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	**87.92**	81.76	**84.73**	**84.04**	79.18	**81.54**
Roberta-100	86.14	**82.20**	84.12	81.34	**79.88**	80.60
ERNIE-100	87.00	80.47	83.61	83.73	76.39	79.90

Among the three Chinese pre-training models under People's Daily corpus, Chinese BERT model is better than other models. The retraining effects of the three Chinese pre-training language models on Amazon commodity corpus and restaurant review corpus are shown in Table 7 and Table 8 respectively.

TABLE 7. Results of Re-pre-training Model on the Amazon Commodity Corpus

model	DEV			TEST		
	P	R	F1	P	R	F1
word2vec-100	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	**86.98**	**82.02**	**84.43**	**84.19**	**79.58**	**81.82**
Roberta-100	81.30	81.15	81.23	80.04	77.09	78.54
ERNIE-100	83.17	81.59	82.37	79.98	77.59	78.77

Table 8. Results of Re-pre-training Model on the Restaurant Review Corpus

model	DEV			TEST		
	P	R	F1	P	R	F1
word2vec-100	86.94	80.12	83.40	81.63	76.95	79.22
BERT-100	**87.25**	**82.80**	**84.97**	**83.68**	**80.18**	**81.89**
Roberta-100	84.02	81.33	82.65	80.35	78.59	79.46
ERNIE-100	85.73	80.47	83.01	82.07	75.70	78.76

After retraining under the commodity review corpus and restaurant reviews, the effect of Chinese BERT pre-training was significantly improved, but the sub-accuracy of Roberta and ERNIE models decreased significantly, indicating that a lot of semantic information was lost after retraining.

In order to verify the influence of different size training corpus on the pre-training language model, the experimental results of increasing the training corpus are shown in Table 7.

Table 9. Results of Re-pre-training Model on the Different Sizes of Restaurant Review Corpus

corpus	DEV			TEST		
	P	R	F1	P	R	F1
306MB	87.25	**82.80**	84.97	83.68	**80.18**	**81.89**
616MB	**88.83**	82.45	**85.82**	83.97	79.28	81.56
919MB	87.71	82.02	84.77	**84.10**	78.49	81.20
1.3GB	88.47	81.59	84.89	83.93	76.99	80.31

As shown in Table 9, increasing the training corpus can significantly improve the performance of remote supervised named entity recognition, but with the increase of data, the effect tends to be stable.

3.4. Entity Linkong Experiments

In order to verify the applicability of the model proposed in our paper, we use Wikipedia and computer paper data for entity acquisition in the computer domain. Entity linking adopts two-classification model based on BERT+ graph embedding.

Table 10. Results of Computer Domain Entity Link

model	F1
BERT-Graph	90.34

The results of computer remote supervision of named entity recognition are shown in Table 11.

Table 11. Results of Computer Domain

model	DEV			TEST		
	P	R	F1	P	R	F1
BERT	92.14	90.56	91.34	90.33	89.55	89.94
BERT-sp	95.36	90.78	93.01	93.57	90.23	91.86

As can be seen from Table 11, the BERT model uses Chinese BERT as input, and BERT-sp is the result of re-training using commodity review 616MB training corpus on the basis of Chinese BERT pre-training. The results show that the effect of re-pre-training is increased in all indicators. Therefore, the effect of re-pre-training can improve the performance of the task in the distant supervision of named entity recognition task.

4. Conclusion

Recent work in natural language processing has focused on domain named entity recognition. In this paper, we show a distant supervised named entity recognition, which combines Wikipedia and paper data to obtain labeled corpus, and applies the labeled data to distant supervised named entity recognition. Finally, the method is applied to the computer domain entity recognition, and the experimental results show that the method can adapt to the domain named entity recognition task.

We use the entity graph and the paper to carry on the entity link, and the entity link model is represented by the embedded representation of the text and the entity graph. As semi-structured data, the paper data can also be used to build a graph, and the next step is to link the entity graph with the paper graph. Secondly, in

the graph attention neural network, adding PMI to the graph pruning to obtain the important attributes of the entity can obtain a better entity graph embedding representation to improve the performance of the entity link.

5. References

[1] Yang, Y., Chen, W., Li, Z., and Zhang, M, 'Distanly supervised NER with partial annotation learning and reinforcement learning,' Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 2-26, pp. 2159-2169, 2018.

[2] Cao, Y., Hu, Z., Chua, T. S., Liu, Z., and Ji, H, 'Low-resource name tagging learned with weakly labeled data,' arXiv preprint arXiv:1908.09659.

[3] Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., and Zhang, C, 'Bond: Bert-assisted open-domain named entity recognition with distant supervision,' In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, USA, August 23-27, pp. 1054-1064, 2020.

[4] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,' Proceedings of NAACL-HLT, pp. 4171-4186, 2019.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Polosukhin, I. 'Attention is all you need,' Advances in neural information processing systems, pp. 5998-6008, 2017.

[6] Peters M E, Neumann M, Iyyer M. 'Deep contextualized word representations,' Proceedings of NAACL-HLT, pp. 2227-2237.

[7] Wang, R., Liu, W., and McDonald, C, 'Featureless domain-specific term extraction with minimal labelled data,' Proceedings of the Australasian Language Technology Association Workshop 2016, pp. 103-112, 2016.

[8] Amjadian, E., Inkpen, D., Paribakht, T., and Faez, F, 'Local-global vectors to improve unigram terminology extraction,' Proceedings of the 5th International Workshop on Computational Terminology, pp. 2-11, 2016.

[9] Hätty, A., and im Walde, S. S, 'Fine-grained termhood prediction for german compound terms using neural networks,' Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions, Santa Fe, New Mexico, USA, August 25-26, pp. 62-73, 2018.

[10] Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S, 'Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks,' In Interspeech, pp. 2072-2076, 2018.

[11] Shah, S., and Reddy, S, 'Similarity Driven Unsupervised Learning for Materials Science Terminology Extraction,' Computación y Sistemas, vol. 23. pp. 1005-1013, 2019.

[12] Šajatović, A., Buljan, M., Šnajder, J., and Bašić, B. D. 'Evaluating automatic term extraction methods on individual documents,' Proceedings of the Joint Workshop on Multiword Expressions and WordNet, pp. 149-154, 2019.

[13] Kessler, R., Béchet, N., and Berio, G, 'Extraction of terminology in the field of construction,' 2019 First International Conference on Digital Data Processing, pp 22-26, 2019.

[14] Pollak, S., Repar, A., Martinc, M., and Podpečan, V, 'Karst exploration: extracting terms and definitions from karst domain corpus,' Proceedings of eLex, pp. 934-956, 2019.

[15] Terryn, A. R., Hoste, V., and Lefever, E, 'In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora,' Language Resources and Evaluation, vol.54, pp.384-418, 2020.

[16] Rigouts Terryn, A., Hoste, V., Drouin, P., and Lefever, E, 'Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset,' 6th International Workshop on Computational Terminology, pp. 85-94, 2020.

[17] Kafando, R., Decoupes, R., Valentin, S., Sautot, L., Teisseire, M., and Roche, M, 'ITEXT-BIO: Intelligent Term EXTraction for BIOmedical Analysis,' Health Information Science and Systems, vol. 9, pp. 1-23, 221.

[18] Kipf, T. N., and Welling, M, 'Semi-supervised classification with graph convolutional networks,' arXiv preprint arXiv:1609.02907, 2016.

[19] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y, ‘Graph attention networks,’ arXiv preprint arXiv:1710.10903, 2017. arXiv preprint arXiv:1710.10903.

[20] Zhang, Y., Yu, X., Cui, Z., Wu, S., Wen, Z., and Wang, L, ‘Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks,’ Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 334-339, 2020. arXiv preprint arXiv:2004.13826.

[21] Huang, L., Ma, D., Li, S., Zhang, X., and Wang, H, ‘Text Level Graph Neural Network for Text Classification,’ Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3444-3450, 2019. arXiv preprint arXiv:1910.02356.

[22] Zheng, B., Wen, H., Liang, Y., Duan, N., Che, W., Jiang, D., and Liu, T, ‘Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension,’ Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 678-6718. arXiv preprint arXiv:2005.05806.

[23] Zhan fei, ZHU Yanhui, and LIANG Wentong, ‘Entity Linking Via BERT and TextRank Keyword Extraction,’ Journal of Hunan University of Technology,vol.34, pp. 63-70, 2020.

[24] Luong, M. T., Pham, H., and Manning, C. D, ‘Effective approaches to attention-based neural machine translation,’ ArXiv Preprint ArXiv:1508.04025, 2015.

[25] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A, ‘Don't stop pretraining: adapt language models to domains and tasks,’ arXiv preprint arXiv:2004.10964, 2020.