IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 6, 2020, accepted December 1, 2020, date of publication December 4, 2020,
date of current version December 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3042672

A Neural Relation Extraction Model for Distant Supervision in Counter-Terrorism Scenario

JIAQI HOU[1], XIN LI[1], RONGCHEN ZHU[1], CHONGQIANG ZHU[2],
ZEYU WEI[3], AND CHAO ZHANG[4]

[1]School of Information Technology and Cyber Security, People's Public Security University of China, Beijing 100038, China
[2]Technical Investigation Detachment, Lianyungang Public Security Bureau, Lianyungang 222006, China
[3]Patrol Special Police Brigade of Keqiao, Shaoxing Public Security Bureau, Shaoxing 312000, China
[4]Network Security Detachment, Lianyungang Public Security Bureau, Lianyungang 222000, China

Corresponding author: Xin Li (lixin@ppsuc.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC0803700, in part by the People's Public Security University of China through 2019 Basic Research Operating Expenses New Teacher Research Startup Fund Project under Grant 2019JKF424, and in part by the Natural Science Foundation of China under Grant 41971367.

ABSTRACT Natural language processing (NLP) is the best solution to extensive, unstructured, complex, and diverse network big data for counter-terrorism. Through the text analysis, it is the basis and the most critical step to quickly extract the relationship between the relevant entities pairs in terrorism. Relation extraction lays a foundation for constructing a knowledge graph (KG) of terrorism and provides technical support for intelligence analysis and prediction. This paper takes the distant-supervised relation extraction as the starting point, breaks the limitation of artificial data annotation. Combining the Bidirectional Encoder Representation from Transformers (BERT) pre-training model and the sentence-level attention over multiple instances, we proposed the relation extraction model named BERT-att. Experiments show that our model is more efficient and better than the current leading baseline model over each evaluative metrics. Our model applied to the construction of anti-terrorism knowledge map, it used in regional security risk assessment, terrorist event prediction and other scenarios.

INDEX TERMS BERT, relation extraction, distant supervision, selective attention mechanism, BERT entity encoding.

I. INTRODUCTION

The big data analysis software played an essential role in intelligence analysis in the U.S. government's pursuit of Osama bin Laden; Besides, it also helped banks recover billions of dollars hidden by former Nasdaq chairman Madoff Bernie Madoff. It is hard to do anything today completely avoid the Internet, criminals are no exception; The enemy has used the Internet means, the national anti-terrorism must use big data analysis. Having the ability to analyze data is not a 100% solution, but certainly a significant increase in counter-terrorism capabilities and the cost of committing terrorist acts.

However, the traditional method of manual screening is challenging to satisfy the unstructured form of big data in the network. The unstructured big data refers to the information such as text, image and video, which is much larger than the traditional structured data, and the unstructured semantic accurate search and mining becomes particularly important. NLP technology is the better tool to realize it. NLP is the technology to develop applications or services that can understand human language. NLP tools and techniques, combined with text mining, machine learning, and ontology modelling, have become the first line of defence for military security threat prediction, detection, and early-stage prevention. Today, big data and data technology, through improved collaboration and data analysis, reduce the complexity of the intelligence investigation process so that institutions can more easily detect national security threats.

The characteristic of big data on public security is low-value density and diverse complexity. Therefore, it is critical to combine the knowledge system of public security, integrate all kinds of databases into a unified KG, express knowledge OWL the international unified ontology network language, and use KG technology to realize the reasoning and application of public security big data. The KG lays a theoretical foundation for the construction of a set of semantic knowledge search and mining platform (Figure 1). Among

The associate editor coordinating the review of this manuscript and approving it for publication was Huiling Chen.

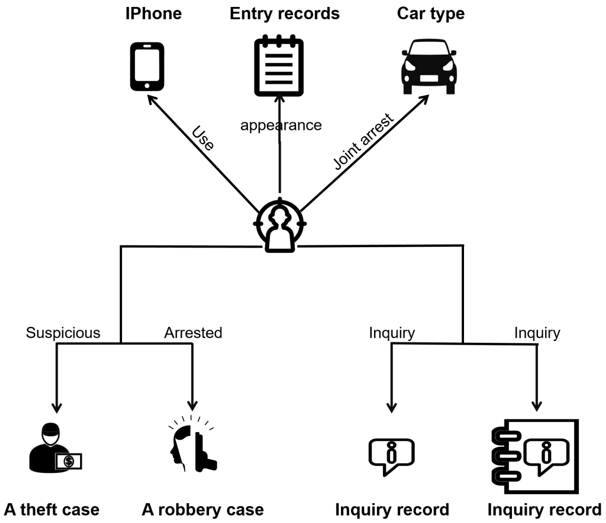


FIGURE 1. Application of counter-terrorism scenario.

them, relation extraction is the crucial step of the construction of anti-terrorism KG for public security. Relation extraction is the process of extracting semantic relations between two or more entities from text. For instance, 'International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911', and we can extract the following relation triples from the above text: Founding-year (IBM, 1911) and Founding-location (IBM, New York). This process is relation extraction.

Traditional relation extraction relies a lot on manual tagging [1]; However, distant supervision can effectively solve this problem. Distant supervision assumes that for a triple (consisting of a pair of entities and a relation) in an existing KG (such as Freebase, Wikipedia), any sentence in the external document library that contains the pair of entities, this relation is reflected some extent. Based on this assumption, the existing knowledge base is corresponding to rich unstructured data, thus generating a large number of training data, thus training a suitable relation extractor. The application of distant supervision to relation extraction can significantly reduce the dependence of manual tagging. More abundant information can be obtained from massive untagged data to improve the accuracy of the classifier [2]. Because of the diversity and complexity of public security data, the establishment of a knowledge graph in the field of anti-terrorism will help the public security to solve the business needs better and even further improve the efficiency of police handling cases through knowledge reasoning. The relation extraction of distant-supervised accords with the current public security data demand and saves the cost of manual marking. At present, the Chinese distant-supervised relation extraction model is still static for the understanding of the text. In this paper, the BERT pre-training model is introduced to dynamically generate word vectors according to the sentences in which the words are located, which can effectively solve the problems of word segmentation ambiguity and synonym ambiguity. Chinese Bert pre-training model consists of a large number of Chinese data training, transmitted generally, the pre-training model generates 100-dimensional word vector representation, and the BERT model can generate 768 dimensions. The understanding of the Chinese text is more multidimensional and deeper. Through the use of BERT model, the BERT embedding, encoding, feature extraction and pooling of public security data are carried out to obtain better sentence vector representation. Then the sentence vector representation is used for the training of the relation extraction model for remote supervision, and finally, better performance is obtained.

In this paper, we propose BERT-att architecture which consists of BERT Model layer, BERT Entity Encoding layer, Bag Attention layer and Classifier layer. First, BERT Model layer carries on the feature extraction to the input sentence and the entity pair. Adoption of pre-training model BERT-base-uncased (12-layer, 768- hidden, 12- heads, 110M parameters) [3], at the Embedding layer, transform features into 768- dimensional word vectors, a position vector and a token vector. Add the three embeddings as intermediate results, The results are fed into the Layer-Normalization layer and dropout layer for processing, and we get the final output and return. The BERT Entity Encoding layer, which uses entity supporting information to splice and output the corresponding hidden. On the Bag Attention layer, the instance in each bag learn the weights, and the bag relation vector representation gives the attention of the correct instance a more significant proportion of the weights. Classifier layer, through the full connection layer and other processing layers to complete the construction of the classifier, using Cross-Entropy calculate loss, select the AdamW optimizer for gradient descent.

The distant-supervised algorithm can label the sentences in the external document library based on a small KG, and put the instances with the same entity pair and relation into the same bag, which is equivalent to the automatic annotation of the sample. Through the bag-level relation extraction, the limitation of manual data tagging in supervised learning is broken. The Global Global Vectors for Word Representation (GloVe) is the pre-training model of present relation extraction, it is a 100-dimensional vector, and the BERT model can generate 768-dimensional vector. The BERT-base-uncased can express more semantic information because it was trained with a vast English corpus. BERT generating dynamic word vectors, we can solve the problem of polysemous words. BERT entity encoding layer adds entity description information to assist the representation of learning entity, thus improving the accuracy. The attention model is used to learn the weight of the instance in each bag, the correct label instance given more attention, and the incorrect one's contribution is lower, thus improving the accuracy of classification.

Here, the contribution of this paper can be summarized as follows. (1) In the related Work, we introduce the related Work of distant-supervised relation extraction method.

(2) In section A of the methodology, we introduce the main structure and principles of the BERT pre-training model, and the core mechanism is the part of Scaled Dot-Product attention; Meanwhile, we propose BERT Entity encoding, use the word vectors corresponding to the unique markers in front of the two entities and splice them together. The final sentence-level relation vector representation of each instance is obtained. The pre-training model based on BERT is helpful to solve the words' polysemy problem and capture the better semantic features in sentences. (3) In Section B, the Bag Attention layer we proposed further improves the extraction effect by giving different sentence attention weights and selectively paying attention to the examples in the bag. (4) In Section C, through Linear layer, softmax layer, dropout layer and full connection layer, the relation extraction model is constructed. (5) The proposed model can obtain more accurate relational representation than current leading models. In order to prove the superiority of this model, we compare it with other relational extraction models of different processing layers and give the experimental results and analysis in the simulation experiment. Finally, the conclusion is drawn and summarized.

II. RELATED WORK

In [4], Mintz proposed the application of distant supervision to open domain relation extraction, which is different from traditional predefined relational categories. Distant-supervised automatically constructs a large number of training data by aligning the knowledge base with unstructured text [5], reduces the dependence of the model on manually annotated data, and enhances the model's cross-domain adaptability [6]. Distant supervision combines the advantages of fully supervised probabilistic classifier calculation and unsupervised free domain relation extraction. The hypothesis using At-Least-One was presented in [7], Riedel in 2010 and predicted in conjunction with the undirected graph model; In [8], Hoffmann proposed a multi-instance learning method for Multi-Instance Learning (MultiR) in 2011, using a probabilistic graph model to select instances (based on multi-instance learning) and to increase overlap related to relation extraction systems; In [9], Surdeanu proposed multi-relational multi-label method in 2012; In [10], Zeng was expanded in 2015 based on full-supervised learning, from fully-supervised to distant-supervised, to select the most effective instance for relation prediction in combination with multi-instance learning and PCNN, and to achieve the best results at present; In [11], Lin introduced an attention mechanism based on the previous Zeng 2015 in 2016, using a sentence-level attention mechanism to assign a weight to each sentence to reduce the problem of information loss; In [12], Jiang addressed the problem of Zeng 2015 from another perspective in 2016, and the multi-relation of entity pairs is considered. In [13], Liu proposed the concept of soft tags, using the correct semantic information in the training phase to correct the label of false tagging and reduce the problem of false tagging in 2017; In [14], Huang proposed the weak supervision relation extraction under the deep residual network in 2017 to solve the noise interference caused by distant supervision and enhance the relation extraction effect. In [15], Ji add description information of entities on a PCNN and Attention basis in 2017; In [16], Lin proposed the relation extraction with multi-lingual attention in 2017; In [17], Yang proposed a word-level attention mechanism to pick out more critical words in 2017. In [18], Qin proposed a deep reinforcement learning framework in 2018 to remove false positive instances from the original training set and reconstruct a pure training (test) dataset to improve the accuracy of relational classification. In [19], Zeng introduced reinforcement learning into the relation extraction model of distant supervision in 2018. In [20], Hu improved distantly-supervised relation extraction with Joint Label Embedding in 2019.

The method of weighted summation is more suitable for the methods of noise removal and error annotation, such as RL (Reinforcement Learning) can reduce the number of error annotations, and generative countermeasure network Generative Adversarial Networks (GANs) can remove the large amount of noise generated by distant supervision. This paper is mainly based on Lin *et al*. The traditional Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) or Long short-term memory (LSTM) feature extraction method is changed into a BERT self-attention mechanism; It is easier to calculate. There is no gradient vanishing problem compared to recurrent; It can see farther information, so it is also more suitable for text than CNN. CNN can be seen far away when it is high level, but CNN abstractions that would have to wait until very high levels, self-attention can do it even at the bottom, this is undoubtedly a huge advantage. This paper replaces the classic combination of word2vec and PCNN with the latest BERT pre-training model, establishes the attention weight k each word in the sentence sequence and other words in the sentence sequence, and linearly weights the information at all times of the sentence sequence. Position information is introduced to improve the accuracy of relation extraction further and achieve the state-of-the-art results.

III. METHODOLOGY

This architecture contains Bert Model layer, Bert Entity Encoding layer, Bag Attention layer and Classifier layers (Figure 2).

1. At the BERT model layer, the model effectively gets the representation of the relation vector, implementing input features embedding, Transformer coding and pooling by the BERT Embedding layer, BERT Encoder layer and BERT Pooler layers respectively. Represented by a unique mark [CLS] of the BERT model as a relation vector, insert a unique mark [$\mathrm{E1}_{\mathrm{start}}$], [$\mathrm{E2}_{\mathrm{start}}$] before the two entities.

2. At the BERT Entity Encoding layer, stitching the hidden vectors of these two special marks as a representation of the relation vectors (i.e. the BERTem model).

3. At the Bag Attention layer, the classification strategy is implemented after obtaining the representation of relational

TABLE 1. The full form of the abbreviation in the text.

Full form	Abbreviation
Natural Language Processing	NLP
knowledge graph	KG
Bidirectional Encoder Representation from Transformers	BERT
Global Vectors for Word Representation	GloVe
Multi-Instance Learning	MIL / MultiR
Reinforcement Learning	RL
Generative Adversarial Networks	GANs
Machine learning	ML
Information Retrieval	IR
Embeddings from Language Model	ELMo
Generative Pre-Training	GPT
Continuous Bag-of-Word Model	CBOW
Rectifiedlinearunit	ReLU
Area Under the Curve	AUC
Precision-Recall plot	P-R plot
Recurrent Neural Networks	RNN
Convolutional Neural Networks	CNN
Long short-term memory	LSTM

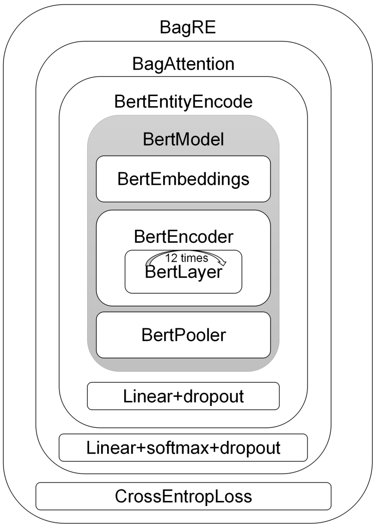


FIGURE 2. Model structure framework.

vectors. A distant-supervised attention strategy is applied to all the instances of relational vectors in a bag to obtain the relational vector representation of the bag [21].

4. On the Classifier layer, the whole model weight is trained by full connection and softmax, and the model weight is used in the attention strategy.

Among them, B donates batch size, L donates the length of sentence, H donates the number of output channels, i.e. how many dimensional hidden vectors are there (here 768) [22].

A. BERT-BASED PRE-TRAINING MODEL

1) BERT MODEL

Like most NLP deep learning models, BERT fed each word (token) in the input text into the embedding layer to convert each word into a vector form $\text{embedss}_{\text{token\_type}}$. Nevertheless, unlike other models, BERT has two more embedding layers to get the other two vectors $\text{embedss}_{\text{words}}$ and $\text{embedss}_{\text{position}}$.

These representations are added by elements to get a synthetic representation. Moreover, this representation is the input to the BERT coding layer (Formula 1).

$$\text{embeds} = \text{embedss}_{\text{words}} + \text{embedss}_{\text{position}} + \text{embedss}_{\text{token\_type}}. \tag{1}$$

All Query, Key, Value come from the same input but are obtained by three different linear matrixes (full connection layers), so they may not be precisely equal.

A self-attention model is to excavate the internal connection of a sentence under the condition that Query, Key, Value three matrices are equal [23]. Calculate the weight of each word in a sentence. And then it weighted to the vector of each word in a sentence. This calculation uses a dot product.

Moreover, the attention in the article is the scaled dot-product method (Formula 2).

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V. \tag{2}$$

The formula QK^T is Query matrix-vector and Key matrix-vector to do dot product. The product of two vector points represents the similarity of the two vectors.

$$sim_{Q,K} = \frac{QK^T}{\sqrt{d_k}}. \tag{3}$$

, d_k = attention head size

To prevent the result of dot product from being too large, a scaling (d_k is the length of Key vector) is made (Formula 3).

$$sim_{scale}(sim_{Q,K,i}) = \frac{\exp(sim_{Q,K,i})}{\sum_j \exp(sim_{Q,K,j})}. \tag{4}$$

The result is normalized by a softmax to a weight of 1 and multiplied to the Value vector (Formula 4 and 5).

$$Attention(Q, K, V) = sim_{scale} \bullet V. \tag{5}$$

Break a sentence of 512 into h, 512/ h for per length. And then each individually weighted attention [24], Q, K, V split separately. Multi-head attention process Q, K, V by h different linear transformation pairs projection. Finally, different attention results are spliced together (Formula 6).

$$texthead_{\text{i}} = Attention(QW_i{}^Q, KW_i{}^K, VW_i{}^V). \tag{6}$$

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions (Formula 7).

$$MultiHd(Q, K, V) = Concat(head_1, \ldots, head_h)W^O. \tag{7}$$

Then take a linear Linear layer (Formula 8), a dropout and a Layer Norm (Formula 9). Multi-head attention is done.

$$MultiHd(Q, K, V) = MultiHd(Q, K, V)A^T + b. \tag{8}$$

$$MultiHd(Q, K, V) = \frac{x - E[MultiHd(Q, K, V)]}{\sqrt{Var[MultiHd(Q, K, V)] + \varepsilon}} * \gamma + \beta. \tag{9}$$

Through a linear Linear layer and activation function, after a linear change, then a Dropout and a normalization, the whole part of the Transformer is completed (Formula 10).

$$M = MultiHd(Q, K, V){A_{M\_Head}}^{T} + b_{M\_head}. \tag{10}$$

A self-selected excitation function is a GeLu function which applies Gaussian error linear element function (Formula 11).

$$\mathrm{M} = \mathrm{GELU(M)} = \mathrm{M}^{*}\Phi\ \mathrm{(M)}. \tag{11}$$

BERT Pooler section is a module used to pool the output of a Bert Encoder, including a Linear layer plus a Tanh() activation function. Applies a linear transformation to the incoming data (Formula 12).

$$P = M{A_{pool}}^{T} + b_{pool}. \tag{12}$$

Tanh() activation use element-wise function (Formula 13).

$$\mathrm{W_{token}} = \tanh(P) = \frac{e^{P} - e^{-P}}{e^{P} + e^{-P}}. \tag{13}$$

2) BERT ENTITY ENCODING

By Entity marker, we insert special marks on both sides of the entity to highlight the entity. In the text $\overrightarrow{\mathrm{x}_i}$, containing an entity pair, insert the unique identification [E1$_{start}$], [E2$_{start}$] into the front of the entity respectively to obtain (Formula 14).

$$\mathrm{x}_i = [x_0 \ldots [E1_{start}]x_i \ldots [E2_{start}]x_j \ldots x_l]. \tag{14}$$

The Sequence output is a matrix of size (batch size,sequence len, hidden state) after the previous BERT Model final output, containing a vector representation of all text $\overrightarrow{\mathrm{x}_i}$ in a batch. Moreover, pos1, pos2 are two matrices of size (batch size,1), representing the starting position of two entities in each context, that is, the corresponding position of [E1$_{start}$], [E2$_{start}$]. Use the unique mark [E1$_{start}$], [E2$_{start}$] corresponding word vector before the two entities and splice it. Finally, each instance is represented by a sentence-level relation vector $\overrightarrow{\mathrm{x}_1}, \overrightarrow{\mathrm{x}_2}, \ldots, \overrightarrow{\mathrm{x}_i}$.

B. BAG ATTENTION

1) DISTANT SUPERVISION

If the unlabeled text contains an entity pair with a specific relation in the KG, it is assumed that the text also describes the same relation. Although a large amount of data can be obtained through this tagging strategy, it will also be as noisy as many noise data because of its assumptions (because the text containing an entity pair does not necessarily describe the corresponding relation).

2) TO SOLVE THE NOISE PROBLEM CAUSED BY DISTANT-SUPERVISED TAGGING

We usually use the method of MIL to put multiple data into a bag, bag where all sentences contain the same entity pairs. There are i examples in a bag in the model diagram, and the sentence $\mathrm{I}_1, \mathrm{I}_2, \ldots, \mathrm{I}_i$ obtains its corresponding relation vector representation $\overrightarrow{\mathrm{x}_1}, \overrightarrow{\mathrm{x}_2}, \ldots, \overrightarrow{\mathrm{x}_i}$ after the above BERT Entity Encoding layer. For the bag relation vector representation, we obtain it by weighted average (Formula 15).

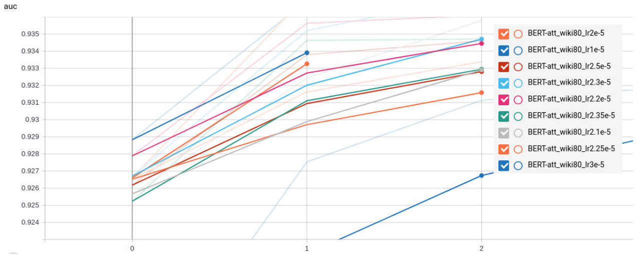


FIGURE 3. The influence of different learning rate models on AUC.

$$\mathrm{s_{bag}} = \sum_{i} \alpha_i \overrightarrow{\mathrm{x}_i}. \tag{15}$$

The vector representation of the bag $\overrightarrow{\mathrm{s_{bag}}}$ is obtained, and then the full connection is added to the softmax classification.

3) SELECTIVE ATTENTION MECHANISMS

Weight α_i is obtained through Attention mechanisms: Let the relation vector of the bag label be expressed as $\overrightarrow{\mathrm{r}}$, for instance I_i in the bag, we calculate the matching degree e_i of its relation vector representation $\overrightarrow{\mathrm{x}_i}$ with the relation vector representation $\overrightarrow{\mathrm{r}}$ of the bag label [25], [26]. For the original PCNN+att, the formula e_i is as follows (Formula 16).

$$e_i = \mathrm{x}_i \mathrm{Ar}. \tag{16}$$

A where the weighted diagonal matrix is. Moreover, here we are going to calculate the dot product between two vectors, as follows (Formula 17).

$$e_i = \mathrm{x}_i \mathrm{r}. \tag{17}$$

By softmax calculating the total proportion of the matching degree e_i of the current relation, we can get the weight α_i (Formula 18).

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_i)}. \tag{18}$$

The attention part of the model is shown below (Figure 3):

Through e_i, that is, softmax get the corresponding weight α_i, multiply and sum it with the relation vector representation $\overrightarrow{\mathrm{x}_i}$ according to the element meaning, that is, carry on the weighted average operation, can obtain the relation vector representation $\overrightarrow{\mathrm{s_{bag}}}$ for the bag.

C. CLASSIFIER LAYER

First, for the vector representation $\overrightarrow{\mathrm{r}}$ of bag label relations, after obtaining the vector representation $\overrightarrow{\mathrm{s_{bag}}}$ of bag, we use full connection plus softmax for classification. The weight matrix obtains the vector representation $\overrightarrow{\mathrm{r}}$ of the bag label relation in the fully-connected layers like the index below the embedding layer.

1) FULL CONNECTION LAYER

Next we use the full connection layer for classification, in which a relational vector representation $\overrightarrow{s_{\text{all\_bags}}}$ of a bag shape (bagN, hidden state) is the weight of the full connection layer $\overrightarrow{R_{\text{fc}}}$ shape (num relation,hidden state), and $\overrightarrow{\text{Output}_{\text{result}}}$ is the classification results. Multiply the elements of $\overrightarrow{R_{\text{fc}}}$ and $\overrightarrow{s_{\text{all\_bags}}}$ one by one and broadcast the results in the last dimension, that is the dot product (Formula 19).

$$\text{Output}_{\text{result}} = \overrightarrow{S_{\text{all\_bags}}} \cdot \overrightarrow{R_{\text{fc}}}. \tag{19}$$

We can see that in the process of classification, for a bag vector representation $\overrightarrow{s_{\text{bag}}}$, that is, a row in $\overrightarrow{s_{\text{all\_bags}}}$, we use it to calculate a dot product with each column of the full connection layer weight $\overrightarrow{R_{\text{fc}}}$. The value obtained by dot product is used as the probability of this bag to describe the corresponding relationship. The form of the dot product is used to calculate the basis of matching degree e_i. For this reason, we can think of each column of the full connection layer weight as a vector representation $\overrightarrow{r}$ of the corresponding relation.

At the time of calculating e_i, we only need to use the form similar to the embedding, and the corresponding relation vector $\overrightarrow{r}$ can be taken out with the $\overrightarrow{x_i}$ to calculated dot product, to obtain the matching degree e_i.

2) SOFTMAX AND DROPOUT LAYER

The output is normalized to filter redundant information by softmax and dropout operations (Formula 20 and 21).

$$\text{Outp}_{\text{result}} = \mathit{soft}\max(\overrightarrow{\text{Outp}_{\text{result}}}). \tag{20}$$

$$\text{Outp}_{\text{result}} = \mathit{dropout}(\overrightarrow{\text{Outp}_{\text{result}}}). \tag{21}$$

IV. EXPERIMENTS AND RESULTS ANALYSIS

A. DATASET

1) DATASET SOURCES

Judging from the original text on the OpenNRE, Wiki80 is derived from the data set FewRel released by Tsinghua.

2) INTRODUCTION OF DATASETS

This data set contains 80 kinds of relations; The number of each relation is 700, a total of 56000 samples. There is 56000 vital and train together.

Wiki80 folder contains three files: The comparison table of Wiki80_rel2id.json include relations and their indexes, a total of 80 relations, is different from that in the Semeval, which does not contain relation between entities. Wiki80_train.txt & wiki80_val.txt include trian (50400 samples), val (5600 samples), total 56000 samples. Data sets do not contain test sets. The format of the sample is similar that in the Semeval, but id attribute is added to head and tail entities. Wiki80 data set uses artificial precision, does not contain noise.

B. EXPERIMENTAL SETTINGS

The experimental environment is as shown below: Ubuntu 20 LTS, 32 G memory, python 3.6, PyTorch 1.6.0 and NVIDIA 1080Ti graphics. 1. The initial learning rate of the optimizer is set to 2e-5, the weight decay is set to 1e-5 using the optimization method. 2. Meanwhile maximum sentence length (max length) for each sentence is set to 86 words. 3. Addition, the batch size is 64 (batch size), and the program is trained to 100 epoch. 4. Embedding vector dimension is 768 dimensions. 5. AdamW is chosen as the optimizer.

Parameter Settings is as follows (Table 2):

TABLE 2. Hyperparameter settings.

Hyperparameter name	Hyperparameter setting
batch_size	64
optim	adamw
Learning rate	2e-5
weight_decay	1e-05
max_length	86
max_epoch	100
attention_probs_dropout_prob	0.1
hidden_act	gelu
hidden_dropout_prob	0.1
num_attention_heads	12

C. EVALUATION METRICS

In the ML (Machine learning), NLP, IR (Information Retrieval) and other fields, evaluation is a necessary job, and the standard evaluation indicators are as follows: Accuracy, Precision, Recall and F1-Measure.

Accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set. That is the accuracy of the test data set when the loss function is 0-1 loss (Formula 22).

$$acc = \frac{\sum_{c=1}^{C} TP_c}{N}. \tag{22}$$

The formula for Recall is that it calculates the proportion of all "correctly retrieved item (TP)" to all" item (TP + FN)" that should be retrieved (Formula 23).

$$R_c = \frac{TP_c}{TP_c + FN_c}. \tag{23}$$

The formula for Precision is that it calculates the proportion of all "correctly retrieved item (TP)" to all" actually retrieved (TP + FP)" as a result (Formula 24).

$$P_c = \frac{TP_c}{TP_c + FP_c}. \tag{24}$$

F1 value is the harmonic mean of Precision and Recall (Formula 25).

$$F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{25}$$

D. EXPERIMENTAL RESULT ANALYSIS

The following six experiments are conducted to evaluate our proposed model in this section. The first experiment evaluates the effects of the learning rate on F1 and AUC by training models with different learning rates. In the second experiment, the F1 and AUC of four different training models PCNN-att, CNN-att, BERT-avg, and BERT-att, were

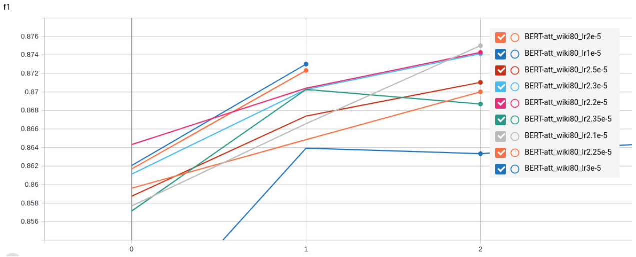


FIGURE 4. The influence of different learning rate models on F1.

TABLE 3. The influence of different learning rate models on AUC and F1.

Learning Rate	1e -5	2e-5	2.1e-5	2.2e-5	2.25e-5	2.3e-5	2.35e-5	2.5e-5	3e-5
AUC	93.18	**93.73**	93.58	93.61	93.34	**93.73**	93.47	93.46	93.70
F1	86.45	87.23	**87.5**	87.43	87	87.41	86.87	87.1	87.30

TABLE 4. F1 and AUC value on different models.

Models	BERT-att	BERT-a vg	PCNN-att	CNN-att
AUC	**93.73**	93.35	84.48	83.86
F1	**87.23**	86.93	77.40	76.33

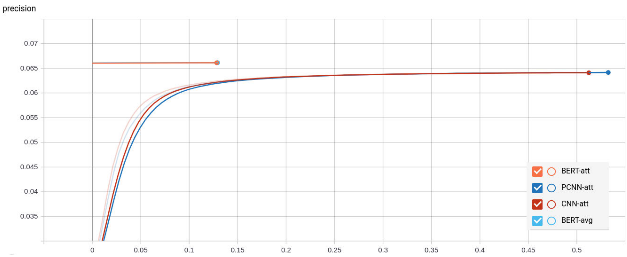


FIGURE 5. Precision value on different models.

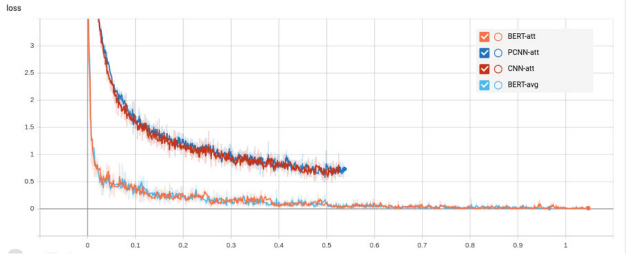


FIGURE 6. Loss convergence comparison.

compared in the same dataset to analyze the performance of different models in F1 and AUC. In the third experiment used the Precision of PCNN-att, CNN-att, BERT-avg, and Bert-att to compare and analyze the influence of different models on Precision. The fourth experiment evaluates the loss convergence effects of the four models are compared. The fifth experiment mainly compares the performance of four different models on the Precision-recall diagram. The sixth experiment evaluates the statistical significance of the proposed model by The Precision at the top N predictions (P@N) [27].

To conclude, we can see from the experiments that our BERT-att achieves state-of-the-art results on F1 values, AUC values, P-R diagram, and P@N.

The learning rate controls the learning progress of the model. If the learning rate is too large, the loss function may directly exceed the global optimal point. At this time, the loss is too large or NaN [28]. The choice of learning rate has a great influence on the model. The following figures (Figure 3 and 4) and tables (Table 3) are the effects of different models on F1, Precision and AUC values when using different learning rates.

For the same dataset trained wiki80, the current leading distant-supervised bag-level relation extraction model PCNN-att's F1 value is 77.40, and its AUC value is 84.48. The F1 of our proposed model is reached 87.23, and its AUC value reached 93.73 (Table 4). From the diagram (Figure 5), it shows that the Precision of our model run faster and higher than other models. BERT has drawn on the Embeddings from Language Model (ELMO), Generative Pre-Training (GPT), and the Continuous Bag-of-Word Model (CBOW) to suggest that the introduction of a priori linguistic knowledge has always been one of the NLP's main objectives, especially in the context of deep learning. However, there has never been the right solution, and the BERT model seems to be a natural and concise solution to this problem, which is the principal value of these methods. BERT pre-training model has a significant effect on the F1 and the AUC. Moreover, our model BERT-att is also better than the BERT-avg model on F1 and AUC. It is visible that the bag-attention layer can improve the model.

In the diagram (Figure 6), the dark blue color line is the PCNN-att loss convergence of the model, and the red color line is the CNN-att loss convergence of the model. And the light blue color line is the BERT-avg loss convergence of the model, the orange color line in the diagram is the convergence loss of our BERT-att model. It shows that our model converges the fastest, and the performance is the best. The other two models need about seventy epochs of training, and our model can converge within four epochs.

Our model's overall performance on the x and y-axis is significantly better than that of the current leading distant supervised bag-level relation extraction model PCNN-at and CNN-att in the Precision-Recall diagram. (Figure 7) Also, our model BERT-att performs better than BERT-avg in the P-R diagram. And it can be seen that the bag attention layer improves the model. One is the need for a more robust feature extractor for the current NLP, where Transformer performance is significantly more potent than the GloVe pre-training model, which PCNN-att and CNN-att used; A word representation tool based on global word frequency statistics (count-based and overall statistics). GloVe using the co-occurrence matrix, at the same time, local information and comprehensive information are considered. Like word2vec, fasttext, the resulting word vectors of GloVe are static, they all can not solve polysemous words problem. Furthermore,

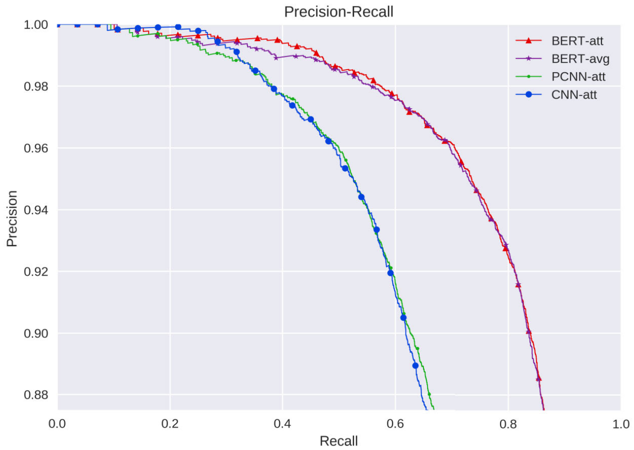


FIGURE 7. Performance of four different models on Precision-recall.

TABLE 5. Performance of four different models on Precision@N.

P@N	1000	2000	3000	Mean
BERT-att	0.9980	**0.9956**	**0.9857**	**0.9931**
BERT-avg	0.9960	0.9920	0.9830	0.9903
PCNN-att	0.9970	0.9840	0.9560	0.9790
CNN-att	**0.9990**	0.9850	0.9533	0.9791

the GloVe loss function is the GloVe's fatal problem, the loss function shows that after adding any constant vector to the GloVe word vector, it is still the loss function's solution. The problem is more prominent if we add a unique constant; Word vectors are very close, So we lose the meaning of word vectors. Therefore, the word vector is trained with GloVe to be checked. At present, it is seen that pre-training BERT is very useful and very concise. It is visible BERT pre-training model has a significant effect on Precision and Recall value.

From the table (Table 5), we can see that the prediction effect of top P@N in our proposed model is better than that of the leading distant-supervised bag-level relation extraction model PCNN-att and CNN-att models. Our model BERT-att performs better than the BERT-avg; It shows that the bag-attention layer influences a lot in top P@N prediction.

The F1, AUC, top P@N and Precision-Recall of our model reach the value of the best distant-supervised bag-level relation extraction model, and our model converges faster the effect is much better than other models. The Bag-attention layer can improve the model, add weight information, and retain more valuable data. Our model as a whole performs better on the x and y axes than the current leading model.

V. CONCLUSION SUMMARY AND FUTURE WORKS

We proposed a neural relation extraction model for distant supervision in this paper, which is named Bert-att. The application of the BERT pre-training model to the relation extraction of distant supervision significantly improves the extraction accuracy and efficiency; BERT is the latest and most advanced pre-training model. BERT pre-training and fine-tuning can solve 11 NLP tasks. The Transformer used are more efficient than RNN, CNN, LSTM and can capture the longer distance dependencies. Compared with the previous pre-training model, and it captures bidirectional context information in the real sense. The previous models are vector representation by word2Vec pre-training models, each word will generate 100-dimensional word vectors statically, but BERT pre-training model can dynamically generate 768-dimensional word vector representation. At the same time, the problem of polysemy and syncopation ambiguity is solved, which dramatically improves the accuracy of each word expression and the accuracy of relation extraction. From experiments, we can see that our model is improved and outperforms the leading PCNN-att and CNN-att models in the performance of top P@N, Precision, F1, and AUC. Bag-attention processing performs better than the bag-averaged to obtain more information about useful instances in the package to express its implicit relational information better. The proposed model's performance is better than the current leading PCNN-att and CNN-att models in the diagrams. Our model's loss convergence effect is very prominent, the convergence is the fastest, and the loss value is the lowest. Our model achieves satisfactory performance, the fastest convergence, and the strongest robustness. This paper provides a new idea for public security analysis in the field of anti-terrorism network big data. In the future, we will try to apply more advanced NLP models to text analysis in the field of counter-terrorism and develop more and more robust knowledge extraction models.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful feedback and suggestions.

REFERENCES

[1] J. Hou, X. Li, H. Yao, H. Sun, T. Mai, and R. J. I. A. Zhu, "Bert-based chinese relation extraction for public security," *IEEE Access*, vol. 8, pp. 132367–132375, 2020.

[2] J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, *Natural Language Processing and Chinese Computing*, vol. 11838. Cham, Switzerland: Springer, Oct. 2019.

[3] X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, *Natural Language Processing and Chinese Computing*, vol. 10619. Cham, Switzerland: Springer, 2018.

[4] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int.*, 2009, pp. 1003–1011.

[5] E. Haihong, X. Zhou, and M. Song, "Distant supervised relation extraction based on recurrent convolutional piecewise neural network," in *Proc. Int. Symp. Signal Process. Syst. (SSPS)*, 2019, pp. 169–175.

[6] Z. Li, Y. Sun, J. Zhu, S. Tang, C. Zhang, and H. Ma, "Improve relation extraction with dual attention-guided graph convolutional networks," *Neural Comput. Appl.*, pp. 1–12, Jun. 2020.

[7] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2011, pp. 148–163.

[8] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, pp. 541–550.

[9] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 455–465.

[10] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1753–1762.

[11] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2124–2133.

[12] L. Ouyang, H. Tang, and G. Xiao, "Chinese text relation extraction with multi-instance multi-label BLSTM neural networks," in *Proc. 31st Int. Conf. Softw. Eng. Knowl. Eng.*, Jul. 2019, pp. 1471–1480.

[13] T. Liu, K. Wang, B. Chang, and Z. Sui, "A soft-label method for noise-tolerant distantly supervised relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1790–1795.

[14] Y. Huang and W. Y. Wang, "Deep residual learning for weakly-supervised relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–4.

[15] H. She, B. Wu, B. Wang, and R. Chi, "Distant supervision for relation extraction with hierarchical attention and entity descriptions," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[16] Y. Lin, Z. Liu, and M. Sun, "Neural relation extraction with multi-lingual attention," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 34–43.

[17] L. Yang, T. Lok, J. Ng, C. Mooney, and R. Dong, "Multi-level attention-based neural networks for distant supervised relation extraction," in *25th Irish Conf. Artif. Intell. Cogn. Sci.*, Dublin, Ireland, Dec. 2017, pp. 7–8.

[18] P. Qin, W. Xu, and W. Y. Wang, "Robust distant supervision relation extraction via deep reinforcement learning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–5.

[19] X. Zeng, S. He, K. Liu, and J. Zhao, "Large scaled relation extraction with reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5658–5665.

[20] L. Hu, L. Zhang, C. Shi, L. Nie, W. Guan, and C. Yang, "Improving distantly-supervised relation extraction with joint label embedding," in *Proc. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3812–3820.

[21] D. Zhao, J. Wang, Y. Zhang, X. Wang, H. Lin, and Z. Yang, "Incorporating representation learning and multihead attention to improve biomedical cross-sentence n-ary relation extraction," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–17, Dec. 2020.

[22] J. Shao, M. L. Yiu, M. Toyoda, D. Zhang, W. Wang, and B. Cui, *Web and Big Data: Third International Joint Conference, AP Web-WAIM 2019, Chengdu, China, August 1–3, 2019, Proceedings, Part I*, vol. 11641. Cham, Switzerland: Springer, 2019.

[23] R. Brochier, "Representation learning for recommender systems with application to the scientific literature," in *Proc. Companion Proc. World Wide Web Conf.*, May 2019, pp. 12–16.

[24] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1, 2018.

[25] Q. Zhang, M. Chen, and L. Liu, "A review on entity relation extraction," in *Proc. 2nd Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Dec. 2017, pp. 178–183.

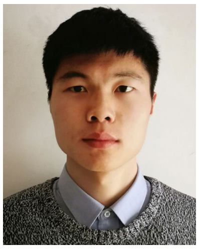
[26] X. Zhu, B. Qin, M. Liu, and L. Qian, *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference, CCKS 2019, Hangzhou, China, August 24-27, 2019*, vol. 1134. Cham, Switzerland: Springer, Aug. 2019.

[27] S. Li, M.-C. Lee, and C.-M. Pun, "Complex zernike moments features for shape-based image retrieval," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 39, no. 1, pp. 227–237, Jan. 2009.

[28] G. Li, J. Sun, and X. Guo, "An ICT system fault analysis technology based on text classification and image recognition," in *Proc. 5th Int. Conf. Comput. Commun. Syst. (ICCCS)*, May 2020, pp. 210–214.

JIAQI HOU received the bachelor's degree in cyber security and law enforcement from the People's Public Security University of China, where she is currently pursuing the master's degree in cyberspace security law enforcement technology. She has published some academic articles and participated in some projects. Her research interests include relation extraction, natural language processing, and knowledge graph.

XIN LI received the Ph.D. degree from the Department of Computer Science, Zhejiang University, in 2007. He is currently an Associate Professor with the School of Information Technology and Cyber Security, People's Public Security University of China, Beijing, China. He has published more than 30 papers in prestigious peer-reviewed journals and conferences. His research interests include cyber security, big data, and artificial intelligence.



RONGCHEN ZHU received the bachelor's degree in cyber security and law enforcement from the People's Public Security University of China, where he is currently pursuing the master's degree in cyberspace security law enforcement technology. He has published some academic articles and participated in some projects. His research interests include risk analysis and assessment, Bayesian network methods and applications, and knowledge graph.

CHONGQIANG ZHU currently working with Lianyungang Public Security Bureau. He has about 30 years of rich police experience in the field of public security. His research interests include forensic technology and countermeasures against crime based on communication and networks.

ZEYU WEI was the Deputy Leader of the Economic Investigation Team, County Public Security Bureau, a Political Instructor of the Sunduan Police Station and the Qianqing Police Station, County Public Security Bureau, the Director of the Anchang Police Station, County Public Security Bureau, a member of the Party Committee, Anchang Town, County Public Security Bureau, and other positions. He is currently working as a Leader with the Patrol Special Police Brigade of Keqiao, Shaoxing Public Security Bureau. His research interests include handling of anti-terrorism and mass incidents.

CHAO ZHANG currently working as a Policeman with the Case Investigation Team, Network Security Detachment, Lianyungang Public Security Bureau. He has years of rich police experience in the field of public security. His research interests include economic crime and cybercrime in the case study.