

# A Named Entity Recognition Model Based on Multi-Task Learning and Cascading Pointer Network

Chaoyang Geng

School of Computer Science and Engineering  
Xi'an Technological University  
Xi'an, China  
E-mail: 541211200@qq.com

Yi Li

School of Computer Science and Engineering  
Xi'an Technological University  
Xi'an, China  
E-mail: 168565012@qq.com

Peng Liu

School of Computer Science and Engineering  
Xi'an Technological University  
Xi'an, China  
E-mail: 643973010@qq.com

Jiejie Zhao

School of Computer Science and Engineering  
Xi'an Technological University  
Xi'an, China  
E-mail: 1838481537@qq.com

**Abstract**—Because, the existing named entity recognition models lack the Specificity of the field, and most of them combine the prediction of entity Location and entity category, which results in the accumulation of errors. So, named entity recognition model based on multi-task learning and pointer network is proposed, and innovations are made in the task Construction and domain entity information utilization in the model of Named entity recognition. This model is based on Transformer with multi-Head attention mechanism, and decomposes traditional tasks in entity recognition tasks and entity classification tasks, and carries out multi-tasks Task learning to reduce the accumulation of errors between tasks. Model in this paper also uses the similarity calculation based on the Comprehensive description of the entity category in entity classification Task for better pertinence of the domain entity. Experiments are conducted on public datasets and domain datasets to prove the advancement of the Model.

**Keywords**-Named Entity Recognition; Multi-Tasks Task Learning; Transformer

## I. INTRODUCTION

Named Entity Recognition refers to the process of extracting specific words from natural language corpus. Nominated entity recognition tasks also have some problems in entity labeling. Traditional model predictions often use BIO [1] or BIOES labeling methods, which define both entity location and entity category. This means that the prediction of each word by traditional named entity recognition models requires that the categories of named entities and entities be combined. This prediction method has a large problem. If any of the named entity subtags or entity category subtags within an entity predicts errors, the entire entity predicts errors, which can easily lead to the accumulation of errors. Currently, researchers have separated traditional

entity recognition tasks into named entity prediction tasks and entity classification modules. Zheng [2] et al. proposed a named entity recognition algorithm based on Multitask learning. Multitask learning is an integrated learning method [3] [4], which improves multiple tasks by training several tasks at the same time. Based on BiLSTM model [5] [6], the traditional named entity recognition model is divided into two modules: named entity prediction task and entity category classification task. It also uses multitask learning to train, which achieves good results, but there are problems that context information other than named entities is not fully used and comprehensive information of entity categories is not introduced. Ding Yi qi [7] et al. proposed a Chinese named entity perception neural network model based on Multitask learning. The traditional named entity recognition task is divided into named entity perception task and entity classification task, and the loss function is optimized to better identify Chinese entities. However, it has the problem of identifying the beginning and end of an entity with two modules, which leads to inconsistency in training and prediction, and lack of organization in the representation of entity categories. This paper proposes a MTL-NER model (a named entity recognition model based on multi-task learning and cascading pointer network) and an entity labeling method based on How Net [8][9][10] semantics. The traditional named entity recognition task is decomposed into global named entity perception and entity classification. The Recognition calculation method based on entity comprehensive description is introduced into How Net knowledge base to classify entities and improve the recognition effect of the overall model. At the same time, entity category description statements are optimized for specific

domains to improve the accuracy of domain entity recognition.

## II. MTL-NER MODEL

The model structure proposed in this paper is shown in Figure 1 below, the overall model is mainly composed of five layers: data preprocessing layer, shared feature extraction layer, multi task learning layer and output layer. In the data preprocessing stage, a semantic entity annotation method based on how net is proposed, which annotates named entities and categories, and constructs a sample set in sentence units. Combined with how net knowledge base, the comprehensive description of domain entity category is constructed as the input data of the model. In the model construction, the task of entity recognition and entity classification is based on the shared feature extraction layer for text vectorization and feature extraction. The sentence is encoded through the shared feature layer to obtain the feature vector of the sample, and then the domain named entity prediction is carried out.

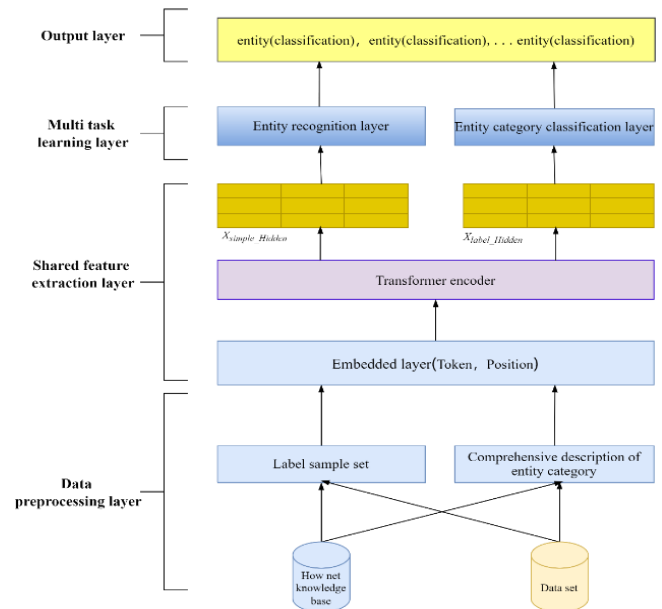


Figure 1. MTL-NER structure diagram

The information of entity classification task comes from the natural language comprehensive description of sample set and entity category, and the same feature extraction layer is used to obtain the feature vector. In the multi task learning layer, the corresponding result vectors are generated from the input feature vectors according to the different tasks. Get the prediction probability results of each task. Finally, the output layer fuses the results of the two tasks, completes the determination of named entities and the classification of entity categories, and obtains the named entity recognition results.

*A. Data preprocessing layer*

In order to reduce the cumulative impact of traditional entity labeling errors, this paper proposes a new label based on How Net semantics, which labels entity categories and domain entities respectively. In How Net, words are composed of one or more semantics, and each semantics is composed of smaller semantic units (semantics) and dozens of dynamic roles. Figure 2 below is an example:

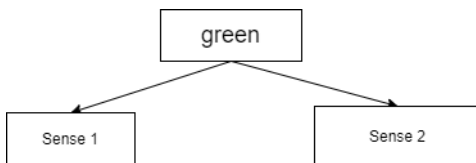


Figure 2. Words and meanings in How Net

The word "green" has two meanings, meaning category 1 is color, and meaning category 2 is environmental protection. Then construct labels to classify by meanings. For the input sequence  $X = \{x_1, x_1, \dots, x_n\}$ , you only need to predict the category, start position, and end position. Based on the above ideas, this topic proposes (Entity category, Start position, End position), taking the domain entity "green" as an example.

Combined with the definition of entity in How Net knowledge base, the meaning item is regarded as the category label of domain entity to improve the accuracy of entity class determination.

Combined with the semantic information of How Net knowledge base, as shown in Figure 4 below, it includes Chinese and English words, semaphores and d DEF\_CONCEPT (combination of semantic and dynamic roles) and attributes; Relationships mainly include dynamic roles, hierarchical relationships, and other domain entity category description statements, which are supplemented by Wikipedia and the specific situation of the corpus.

The introduction of the comprehensive description information of the entity category assists the entity classification task, and improves the domain pertinence and classification accuracy of the model. Through the introduction of the entity category description, the ability of the model to obtain the domain information can be improved, so as to strengthen the pertinence of the specific domain. The example of the description statement constructed in this paper is shown in Table 1 below.

TABLE I. EXAMPLE OF COMPREHENSIVE DESCRIPTION OF DOMAIN ENTITY CATEGORIES

Entity Category	Comprehensive description of entity category
color	Yellow green blue, environment-friendly color, hue, lightness, saturation and various phenomena of light
Environmenta l-friendly	Characteristic value, protection, positive evaluation, low carbon, energy conservation and emission reduction, life, agriculture, circular economy, wind and solar power generation
O	General text

### B. Shared feature extraction layer

In the shared feature extraction layer, named entity recognition tasks and category classification tasks share an embedding layer and feature extraction layer for joint training. According to the sequence length  $l$ , the text  $X = \{x_1, x_2, \dots, x_l\}$  is divided and embedded to obtain the input tensor  $X_{input} \in \mathbb{R}^{b \times l \times d}$  where  $b$  is the batch size,  $l$  is the sequence length, and  $d$  is the word embedding dimension. Then, according to formula (1) (2), the sequence characters are encoded according to the linear changes of sin and cos functions, and the position vector of the characters in the sentence is obtained  $X_{pos} \in \mathbb{R}^{b \times l \times d}$ .

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (1)$$

$$PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (2)$$

Where,  $pos$  is the character position,  $i$  is the character vector dimension,  $d_{model}$  is the hidden dimension of the model. Each encoder is composed of two internal layers: multi head attention mechanism and feedforward neural network. Finally, the input vector of the model is obtained by adding the position code and the word embedded elements:

$$X_{embedding} = X_{input} + X_{pos} \quad (3)$$

In the feature extraction layer, the transformer model based on multi head attention mechanism is used to improve the feature extraction ability of context information. For input vector  $X_{embedding}$  decomposes according to formula (4)

to obtain query matrix  $Q$ , key value matrix  $K$  and numerical matrix  $V$ . As the input of Transformer encoder module.

$$Q^{(h)}, K^{(h)}, V^{(h)} = XW_Q^{(h)}, XW_K^{(h)}, XW_V^{(h)} \quad (4)$$

$W_Q$ ,  $W_K$  and  $W_V$  is the weight parameter matrix,  $h \in [1, n]$  is the head index, and the head number  $n$  is the super parameter. The attention operation is carried out according to formula (5), (6) obtaining the correlation between each word and other words in the sentence, each word vector contains the vector information of other words related to the current sentence. The result of specific operation is shown in formula (7):

$$Attention(Q, K, V) = Soft \max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

$$head^{(i)} = Attention(Q^{(i)}, K^{(i)}, V^{(i)}) \quad (6)$$

$$MultiHead(Q, K, V) = Concat[head^{(1)}; \dots; head^{(n)}]W^o \quad (7)$$

Then,  $MultiHead(Q, K, V)$  and  $X_{embedding}$  carries out residual connection to get  $X_{attention}$ , and normalize the calculation to obtain the standard normal distribution, so as to speed up the training and convergence. The full connection layer feedforward neural network in the encoder is based on  $X_{attention}$  is the input, as shown in formula (8) using ReLU as the activation function and performing two linear mappings to complete the expansion and compression of dimensions respectively.

$$FFN(X) = ReLU(XW_1 + b_1)W_2 + b_2 \quad (8)$$

Where,  $W_1$ ,  $W_2$ ,  $b_1$  and  $b_2$  is the corresponding weight matrix and offset. Finally,  $FFN(X)$  and  $X_{embedding}$  performs a residual connection and normalization calculation to obtain the output  $X_{hidden}$ .

In this paper, the Transformer encoder [11][12][13] based on multi head attention mechanism is used as the feature extraction layer, and the encoder module can be superimposed many times. It realizes unsupervised character level learning and representation of input text sequence under the mechanism of position coding and multi head self-attention.

C. Multi task learning layer

In the named entity recognition task, the shared feature extraction layer extracts the long-distance location dependent features of the context, and outputs the sample vector  $X_{sample\_Hidden}$ , which contains the feature information of the sample. The structure of the named entity recognition model based on this feature vector is shown in Figure3 below:

The vector representation of the sample word level is obtained by the transformer encoder. In this paper, the cascade pointer network is used to realize the sequence annotation task, that is, two 0/1 sequences are generated by two binary classification networks to determine the start and end boundaries (spans) of entities in the sequence. Each span is determined by a head position pointer (start) and a tail position pointer (end). At the same time, multiple binary classification networks are used for entity recognition.

Each word (token) in the input sequence can be represented as the starting position of an element, and the span composed of any two tokens can be represented as any entity, which solves the problem of nested entity and multi class entity recognition, as shown in Figure 4:

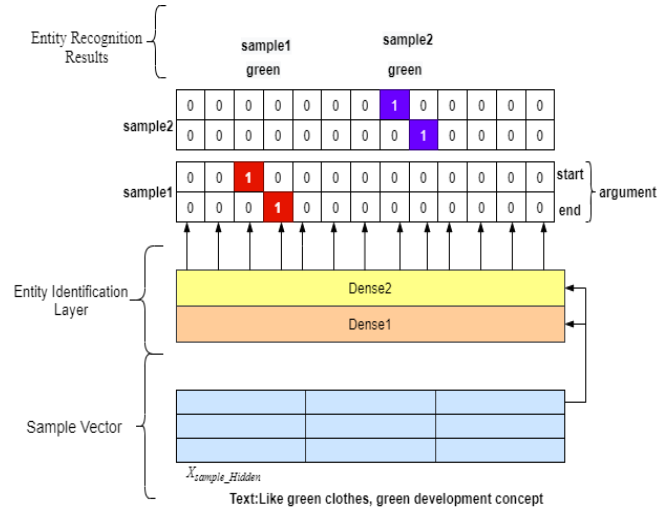


Figure 3. Structure diagram of entity recognition model

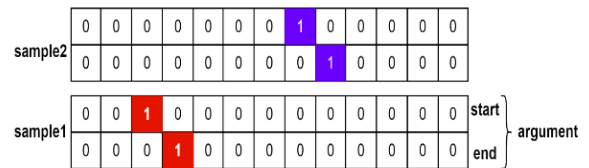


Figure 4. Example of entity recognition

Figure 4 shows the annotation examples corresponding to the input samples. Each entity corresponds to a set of pointer vectors (start, end). By combining the start and end pointer vectors of all entity labels, two-dimensional matrices can be obtained, which are recorded as  $S_s$ ,  $S_e$ . That is, each line in  $S_s$  and  $S_e$  represents an entity type and each column corresponds to a token in the sequence.

In this paper, multiple groups of binary classification networks are used to predict the

possibility that the input sequence is 0/1 at all positions of the start and end pointer vectors corresponding to each entity to determine the start and end positions of the elements. The whole task can be regarded as multi label classification of each token in the input sequence. The probabilities that the  $i$  token is predicted as the starting and ending positions of the elements of entity  $r$  are  $P_i^{s-r}$  and  $P_i^{e-r}$  respectively, as shown in formulas (9) and (10):

$$P_i^{s-r} = \sigma(W_s^r x_i + b_s^r) \quad (9)$$

$$P_i^{e-r} = \sigma(W_e^r x_i + b_e^r) \quad (10)$$

Where,  $x_i = X_{sample\_Hidden[i]}$ , that is, the vector representation of the  $i$  token in the input sequence after passing through the encoder, and the superscript  $s$  and  $e$  are represented as start and end,  $W_s^r$ ,  $W_e^r$  is a trainable weight vector,  $b_s^r$ ,  $b_e^r$  is offset,  $\sigma$  Sigmoid activation letter.

#### D. Entity classification layer

With the help of How Net knowledge base, this paper improves the construction of domain entity classification and entity comprehensive description, and proposes a similarity calculation model based on entity comprehensive description to output the probability of entity category. The structure of entity classification model is shown in Figure 5.

For the input sample eigenvector  $X_{sample\_Hidden}$  and entity category comprehensive description information  $X_{label\_Hidden}$ . This article

first introduces  $X_{label\_Hidden}$  input the full connection layer to realize sentence vector mapping. The formula is as follows:

$$X_{sl} = View(X_{label\_Hidden})W_{label} + b_{label} \quad (11)$$

Where,  $W_{label} \in \mathbb{R}^{(N+1) \times l}$ ,  $N$  is the entity category of the sample set. For the sample set with  $N$  categories, it is necessary to build  $N+1$  entity category description statements. View refers to the method of reconstructing tensor dimension matrix transformation, transforming the vector with dimension  $(N+1) \times d_{label} \times h$  into the dimension of  $h \times (N+1) \times d_{label}$ , and obtaining the entity

category description vector  $X_{sl} \in \mathbb{R}^{h \times (N+1)}$ . Then the vector dot product is used to calculate the similarity between the sample vector and the  $N+1$  entity category description vector. The calculation formula of entity classification is as follows:

$$C_{x1,x2,\dots,xd} = softmax(X_{ss} \bullet X_{sl}) \quad (12)$$

Finally, the output results of category probability corresponding to each input character are obtained respectively.

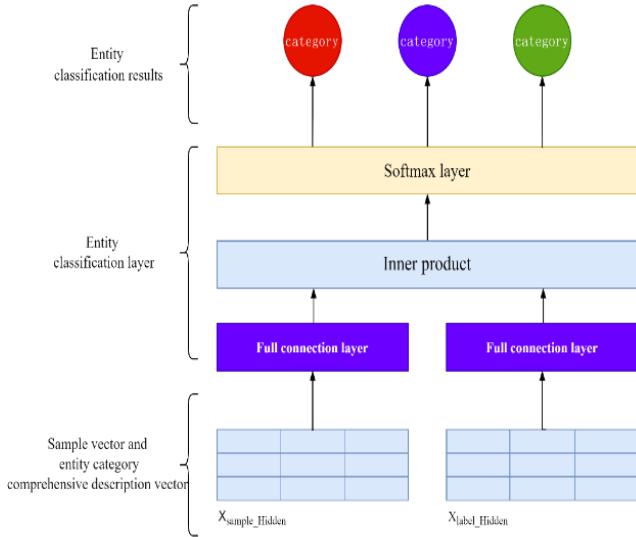


Figure 5. Entity classification model

### E. Output layer and loss function calculation

In this paper, the entity recognition probability and entity classification probability are integrated, and the output results are obtained. The start and end positions of entities in the text are marked by the ruler taking method of double pointers, and the entity categories are marked by multiple binary classification networks. Finally, the loss function formula of the entity recognition part is obtained as follows:

$$loss_{sample} = -\sum_{i=0}^n y_i^{s-r} \log P_i^{s-r} - \sum_{j=0}^n y_j^{s-r} \log P_j^{e-r} \quad (13)$$

Where  $n$  is the length of the input sequence;  $y_i^{(s_r)}$ ,  $y_j^{s-r}$  is a known correct classification label.

In this paper, the loss function calculation formula based on the entity part of character classification result  $C_{X1, X2, \dots, Xd}$  is as follows:

$$loss_{classification} = -\sum_{i=1}^N C_{x_i} \log(Y_{x_i}) \quad (14)$$

Where,  $Y_{x_i}$  entity category label for each character.

Based on the idea of multi task learning, this paper takes the loss function weighting of domain entity recognition task and entity classification as the overall loss function of the model, and the specific formula is as follows:

$$loss = \beta \cdot loss_{sample} + (1 - \beta) \cdot loss_{classification} \quad (15)$$

Among them,  $\beta \in [0, 1]$  is the super parameter of the model, which converges to the optimal value with training to improve the effect of the model. In order to verify the versatility of the model on different data sets, this paper selects four Chinese named entity public data sets as experimental objects. The data sets include MSRA, ontonotes4.0, cluner2020 and CMeEE, and are compared on four Chinese public data sets to verify the progressiveness of the model in this paper.

## III. EXPERIMENT AND A NALYSIS

### A. Data set introduction

MSRA is a Chinese dataset released by Microsoft Research Asia, which comes from the news field and is also the benchmark dataset for Chinese named entity recognition. It contains about 90000 Chinese named entities and annotation data. Entity categories include three categories: location, organization and personnel.

Ontonotes4.0 is a Chinese dataset covering multiple data sources. Sources are not limited to telephone conversations, news agencies, radio news, radio conversations and blogs. This article selects four categories of entities in the version, such as people and organizations.

CLUENER2020 data set is a Chinese fine-grained named entity recognition data set, which is based on the open-source text classification data set THUCNEWS, and selects some data for fine-grained annotation. The data set is divided into 10 different categories and 12000 sentences.

The CMeEE dataset originated from chip 2020 (China health information processing Conference). By extracting entities from sentences, they are classified into nine categories: diseases, clinical manifestations, drugs, medical equipment, medical procedures, body, physical examination, microorganisms and departments, with 25000 sentences.

### B. Experimental indicators and parameter settings

This paper selects the criteria commonly used in named entity recognition tasks: precision (P), recall (R) and F1 score to evaluate the performance of the model, and selects the result of F1 as the main criterion. The specific calculation formulas of accuracy rate, recall rate and F1 score are as follows:

$$precision = \frac{TP}{FP + TP} \quad (16)$$

$$Recall = \frac{TP}{FN + TP} \quad (17)$$

$$F1 = \frac{2 \times precision \times Recall}{precision + Recall} \quad (18)$$

During the experiment, the parameters are shown in Table 2 below:

TABLE II. EXPERIMENTAL PARAMETERS

Parameter	Value
Optimizer	SGD
Learning rate	5e-6
Activate function	ReLU
Entity category length limit	16
Enter length limit	128
Batch size	6
Deep learning framework	Pytorch
Number of GPUs	1

### C. Experimental results and analysis

In this paper, the proposed named entity recognition model MTL-NER based on multi task learning and cascading pointer network is compared with the leading-edge model on the test set of each data set. In the experiment, this paper uses a new ternary tag [entity category, entity start position, end position], while BiLSTM-CRF and BERT-BiLSTM-CRF are trained with traditional BIOES tags, and BERT-MRC uses [entity start position, End position] as a label. In this paper, BiLSTM-CRF is used as the baseline of the experiment. BERT-BiLSTM-CRF is a Transformer based pre training model, which regards the named entity recognition task as a sequence marking task. BERT-MRC<sup>[14][15]</sup> is a retraining Bert model based on machine reading comprehension (MRC). MTL-NER, a named entity recognition model based on multi task learning and cascading pointer network, achieves the best accuracy, recall and F1 on four Chinese named entity datasets. The F1 of this model is improved by 0.77%, 2.62%, 2.27% and 3.32% respectively compared with the model with the best experimental results on the four Chinese entity public data sets of MSRA, OntoNotes4.0, CLUENER2020 and CMeEE. Compared with the baseline model based on BiLSTM-CRF, it is improved by 16.09%, 35.63%, 22.85% and 24.93% respectively, which is enough to show the



progressiveness of this model. The specific performance of each model is shown in the following table 3:

TABLE III. MSRA DATASET MODEL INDICATORS

MSRA			
Model	P (%)	R (%)	F1 (%)
<b>BiLSTM-C RF</b>	87.47	85.23	83.34
<b>BERT-BiLS TM-CRF</b>	95.15	94.85	95.00
<b>BERT-MR C</b>	96.28	95.74	96.01
<b>MTL-NER</b>	97.07	96.43	96.75

TABLE IV. ONTONOTES4.0 DATASET MODEL INDICATORS

OntoNotes4.0			
Model	P (%)	R (%)	F1 (%)
<b>BiLSTM-C RF</b>	73.45	60.07	61.71
<b>BERT-BiLS TM-CRF</b>	79.23	79.58	79.40
<b>BERT-MR C</b>	82.49	81.23	81.56
<b>MTL-NER</b>	84.87	82.56	83.70

TABLE V. CLUENER2020 DATASET MODEL INDICATORS

CLUENER2020			
Model	P (%)	R (%)	F1 (%)
<b>BiLSTM-C RF</b>	67.23	65.42	66.31
<b>BERT-BiLS TM-CRF</b>	77.42	78.15	77.78
<b>BERT-MRC</b>	79.04	80.26	79.65
<b>MTL-NER</b>	82.14	80.79	81.46

TABLE VI. CMEEE DATASET MODEL INDICATORS

CMeee			
Model	P (%)	R (%)	F1 (%)
<b>BiLSTM-CR F</b>	56.41	49.52	52.74
<b>BERT-BiLS TM-CRF</b>	68.98	66.25	67.59
<b>BERT-MRC</b>	71.26	69.34	70.29
<b>MTL-NER</b>	73.13	70.68	71.89

Because more features are integrated into the word vector generation stage, it can quickly achieve better performance in the training process. However, in order to obtain better results on the test set, it is very important to select the appropriate Dropout, which can prevent the model from over fitting and is more robust. In order to determine the appropriate value of Dropout, this paper conducts several groups of comparative experiments on four data sets in the process of selecting Dropout for named entity recognition, F1 value is mainly used as the measurement standard in the experiment, and the results are shown in the following figure6, figure7, figure8 and figure9:

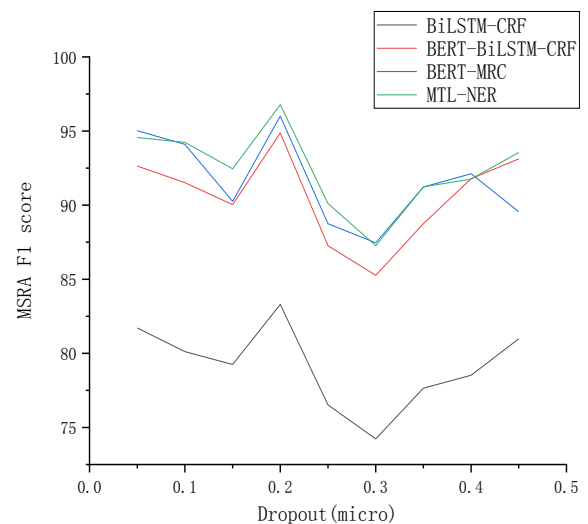


Figure 6. MSRA dataset dropout

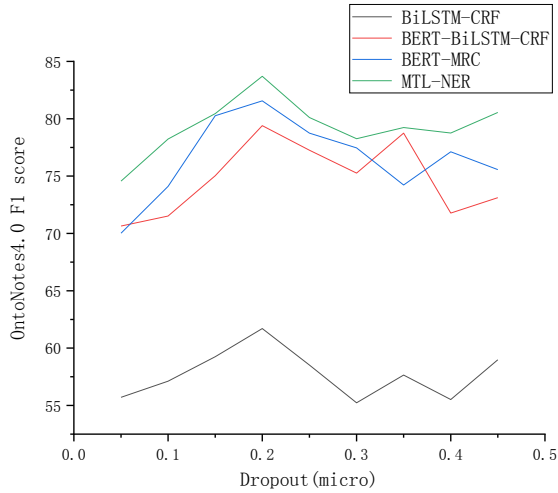


Figure 7. OntoNotes4.0 dataset dropout

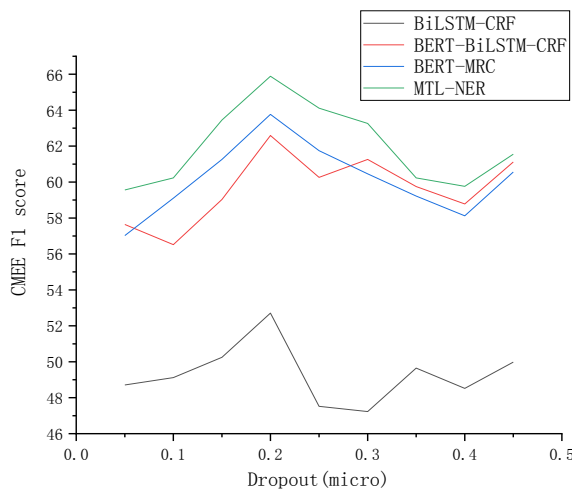


Figure 8. CLUENER2020 dataset dropout

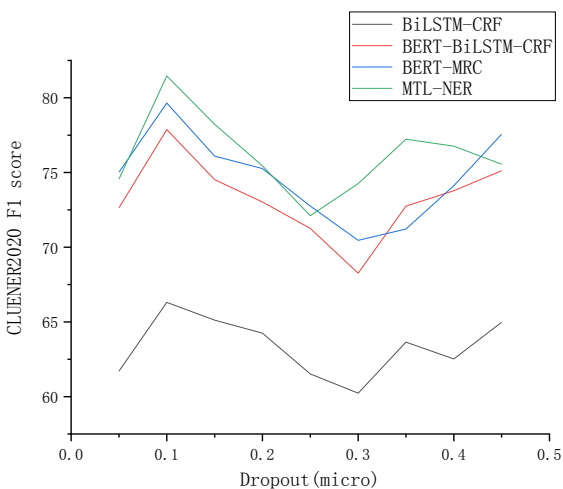


Figure 9. CMeEE dataset dropout

After comparison, in the CLUENER2020 data set, dropout=0.1 is selected, and in MSRA, Onto-Notes4.0 and CMeEE data sets, if dropout=0.2 is selected, the model can achieve the most ideal experimental results. We take the data set of CLUENER2020 data set as an example, and shows the recognition results of domain entities of each model we intercepted:

It is obvious that BiLSTM-CRF model and BERT-BiLSTM-CRF model have poor recognition effect on long text, while BERT-MRC model has multiple recognition errors, and the MTL-NER model proposed in this paper achieves ideal effect on long-span entity recognition.

#### IV. CONCLUDING REMARKS

In the experiment of named entity recognition model, compared with the advanced model, the MTL-NER model in this paper performs well on four Chinese public data sets, and achieves 96.75% F1 on MSRA data set, which is 0.77% higher than the existing advanced model BERT-MRC. This model performs well on multiple data sets covering multiple fields, which proves the versatility of this model for different fields. The experiment also finds that the model based on Transformer is better than the model based on BiLSTM, which verifies the information extraction and utilization ability of Transformer and the rationality of using Transformer as feature extractor in this paper. Finally, experiment Prove the progressiveness of the innovation model in this paper.

#### REFERENCE

- [1] RAMSHAW L A, MARCUS M P. Text Chunking Using Transformation-Based Learning [J]. Springer Netherlands, 1999, 15-176
- [2] Zheng C, Cai Y., Xu J, et al. A boundary-aware neural model for nested named entity recognition[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

- and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).2019:357-366
- [3] RUDER S. An overview of multi-task learning in deep neural networks [J]. arXiv pre-print arXiv:1706.05098, 2017.
- [4] Liu X, He P, Chen W, et al. Multi-Task Deep Neural Networks for Natural Language Understanding[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:4487-4496.
- [5] CHIU J, NICHOLS E. Named Entity Recognition with Bidirectional LSTM-CNNs[EB/OL]. <https://arxiv.org/pdf/1511.08308.pdf>, 2021-02-22.
- [6] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Australia, 2018:1554-1564.
- [7] Ding Yi qi, Yan Dan feng. A neural network model of Chinese named entity boundary perception based on multi task learning [EB / OL], Beijing: China Science and technology papers online [2021-02-22].
- [8] Dong Zhen-dong, Dong Qiang, Ebrary Inc. How Net and the computation of meaning [M] . World Scientific, 2006.
- [9] Dai Liu-ling, Liu Bin, Xia Yu Ning, et al. Measuring semantic similarity between words using How Net [C] / /International Conference on Computer Science & Information Technology, IEEE Computer Society, 2008.
- [10] Ma Yong-qi, Han De-pei, Meng Li-rong, et al. Lexical semantic similarity algorithm based on How-net [J]. Computer Engineering, 2018, 44(6):151-155.
- [11] LI B, KANG X D, ZHANG H L, et al. Named entity recognition in Chinese electronic medical records using transformer-CRF[J].Computer Engineering and Applications, 2020, 56(5):153-159.
- [12] YANG P, DONG W Y. Chinese named entity recognition method based on BERT embedding [J]. Computer Engineering, 2020, 46(4):40-45.
- [13] DONG Z, SHAO R Q, CHEN Y L, et al. Named Entity Recognition in Food Field Based on BERT and Adversarial Training [J]. Computer Science, 2021, 48(5):247-253.
- [14] Li X, Feng J, Meng Y, et al. A unified MRC framework for named entity recognition [J]. arXiv preprint arXiv:1910.11476, 2019.
- [15] Li X, Yin F, Sun Z, et al. Entity-relation extraction as multi-turn question answering [J]. arXiv preprint arXiv:1905.05529, 2019.