Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 7, No. 3, 2023

MFF-CNER: A Multi-feature Fusion Model for Chinese Named Entity Recognition in Finance Securities

Yueping Zhi[1], Xiangxing Tao[1], Yanting Ji[1, *]

[1] Department of Mathematics, Zhejiang University of Science and Technology, Hangzhou 310023, China

[*] Corresponding author: Yanting Ji (Email: yypingzhi@gmail.com)

Abstract: The objective of Chinese financial securities named entity recognition is to extract relevant entities from unstructured Chinese text, such as news, announcements, and research reports, that impact security prices. Recognizing entities in this field is challenging due to the abundance of specialized terms, diverse expressions, and the limited feature extraction capabilities of traditional models. To address this, we propose MFF-CNER, a multi-feature fusion model, to improve the effectiveness of Chinese financial securities named entity recognition. MFF-CNER encompasses several key steps. Firstly, it leverages a BERT pre-training model to capture semantic features at the character level. Secondly, a BiLSTM network is utilized to capture contextual features specific to financial securities text. Additionally, we introduce an Iterated Dilated Convolutional Neural Network (IDCNN) to blend, and extract local features, incorporating an Attention mechanism for weighted feature integration. Finally, the predicted sequences are optimized, and decoded using the Conditional Random Field (CRF). To validate the state-of-the-art performance of MFF-CNER in this domain, we compare it with five popular methods on a Chinese financial securities dataset annotated with the BIO labeling scheme. Notably, MFF-CNER demonstrates superior performance while maintaining compatibility among its components. Furthermore, we evaluate the applicability of MFF-CNER in the Chinese financial securities domain by utilizing public datasets from diverse domains, including social media (WEIBO), and news (MSRA). This research holds practical significance for downstream applications, such as constructing financial securities knowledge graphs, and analyzing factors that influence security prices.

Keywords: Chinese named entity recognition, Financial securities, Multi-feature fusion, BERT.

1. Introduction

Identifying entities with specific meanings from unstructured natural language text, and classifying them into predefined categories is a critical task in Natural Language Processing, commonly called Named Entity Recognition (NER) [1]. The performance of NER models has a significant impact on various downstream tasks, including Q&A systems [2, 3], machine translation [4, 5], and knowledge graph construction [6, 7]. However, most current NER research focuses on general domains [8, 9], biomedical applications [10], clinical records [11], material science[12], , and other fields, while paying little attention to NER in the context of Chinese financial securities.

The financial securities domain is a highly specialized field where many vocabulary terms cannot be understood solely based on their literal meaning. Instead, they require a deep understanding of the financial background, and context to comprehend their specific meanings [13]. NER in this domain possesses distinct characteristics compared to general fields. It involves identifying financial entities like company names, stock names, financial terminologies, financial product names, various abbreviations, as well as general entities such as person names, place names, organizations, and time. The challenges encountered in Chinese NER within the financial securities domain can be summarized as follows:

Non-uniform expression of financial entities: Financial organizations often have multiple expressions for their names. For instance, “网易” (NetEase) is an abbreviation of “网易(杭州)网络有限公司” (NetEase (Hangzhou) Company Limited) but is also known as “NTES” or even referred to as “猪厂” (Pig Factory) on the internet.

Difficulty in recognizing financial terminologies: Financial texts frequently contain entity names comprising a mixture of Chinese, English, and numbers. For example, the fund name “易方达中概互联 50ETF” combines different languages, and alphanumeric characters.

Inadequate availability of well-annotated datasets: The scarcity of properly annotated datasets in the financial securities domain poses challenges, leading to insufficient data support for developing fundamental text processing technologies.

With the advancements in deep learning technology, traditional rule-based, and statistical learning-based NER models are being gradually replaced by deep learning models. The current mainstream neural networks, and pre-trained language models have demonstrated impressive results in NER research across various domains, including some progress in NER tasks within the financial domain [14, 15]. However, the financial securities domain presents specific challenges, such as dispersed content, sparse data, and the lack of structured entities. These challenges make it difficult to directly apply existing NER models to Chinese NER in the financial securities domain. Hence, it becomes necessary to explore novel approaches based on text features.

This study aims to tackle the challenges associated with NER in the Chinese financial securities domain, catering to the requirements of diverse downstream applications. In light of this, we introduce a multi-feature fusion model, MFF-CNER, which effectively considers the specific characteristics of the domain-specific text. Empirical findings unequivocally demonstrate the exceptional performance of this hybrid feature model, capable of extracting information at varying levels of granularity, thereby establishing its superiority in the field.

This paper makes the following contributions:

We propose a multi-feature fusion model called MFF-CNER based on the traditional BiLSTM-CRF [16] model, leveraging pre-trained language models. The MFF-CNER model considers the text features, and entity recognition challenges specific to the financial securities domain, enabling the extraction of text features at different granularity levels, and improving entity recognition accuracy. Moreover, the components of the model do not conflict with each other.

Considering the lack of available datasets for the financial securities domain, we collected financial securities text information from multiple sources, and built a domain-specific dataset using domain expertise. We evaluated the performance of different NER models on this dataset.

Experimental results demonstrate that our model achieves the best precision, recall, and F1 score on the constructed dataset, with values of 88.35%, 89.88%, and 89.10%, respectively.

We validated its applicability in the Chinese financial securities domain by applying the MFF-CNER model to publicly available datasets from different domains.

2. Related Work

Chinese NER has broad application prospects, and many scholars have conducted related research. Research methods mainly fall into three categories: rule-based methods using dictionaries, and rules, statistical machine learning, and deep learning methods.

At the outset, NER assignments were largely dependent on dictionaries that were created manually, and customized rule-based techniques that had the capability of assigning pertinent knowledge bases to general or specific domains, formulating rules on syntax-lexical templates, and then scrutinizing texts for strings that satisfy those dictionaries, and rules [13]. This method has high accuracy but low recall. It mainly depends on domain experts constructing rule templates, which incurs high maintenance costs, and is unsuitable for the financial securities domain's large amounts of data, and fast updating speed.

In general domains, statistical-based methods have achieved remarkable results in NER tasks. Therefore some scholars have applied this approach to NER tasks in other financial domains outside the securities industry [17, 18]. This method requires the manual selection of text features, which are then inputted into a machine learning algorithm. Finally, each character in the sequence is classified according to its corresponding label. Although it has improved its effectiveness compared to previous methods, it still requires significant feature engineering, and involves high time, and labor costs. Currently, in statistical machine learning, only Conditional Random Fields (CRF) are widely used as decoders in new models.

Deep learning techniques based on neural networks have become the dominant approach for NER tasks in modern times, thanks to their strong feature extraction capabilities. Many scholars have studied its application in the financial domain. Huang et al. [16] pioneered in integrating BiLSTM, and CRF models for sequence labeling tasks. BiLSTM can effectively capture global contextual features, and CRF can use sentence-level label information. It has been experimentally verified that this model achieved outstanding results in NER tasks, and has gradually become the mainstream model for such tasks. In order to address the issue of inadequate labeled data in the financial domain, Liu et al. [14] put forward a semi-supervised model that leverages BERT, and Bootstrapping techniques. They utilized the pre-trained BERT model to generate word vectors that could capture the semantic relationships between words, and employed BiLSTM-CRF as the model architecture for training the NER task. The addition of the Bootstrapping method further enhanced the performance of the model, and the experimental results indicated that the Bootstrapping method effectively identified more hidden entities in the dataset with inadequate labels. Moreover, the F1 score of the model with Bootstrapping increased by approximately 0.03 compared to the one without it. Liu et al. [15] proposed an entity recognition method in which the input consists of combined character vectors, and Chinese character Wubi shape embeddings. The model architecture was BiLSTM-CRF, and the authors utilized an iterative learning strategy by embedding the label encoding of the sequence trained in the previous round into the input of the current round to improve the model's prediction results continually. The experiment showed that this approach improved the overall performance of the model. Therefore, in Chinese NER in the financial domain, BiLSTM-CRF is still an effective model architecture [13].

T. Yang et al. [19] proposed the IDCNN-BiLSTM-CRF model that combines features from news, and specialized medical texts to capture more granular feature information, achieving good results on a specialized dataset. Compared to general models, IDCNN [20] has a stronger local feature extraction ability without losing information through pooling. Recently, the BERT-IDCNN-BiLSTM-CRF [21-23] multi-feature extraction method has been applied to multiple domains, and has achieved excellent results. This method adds a BERT pre-trained model to the IDCNN-BiLSTM-CRF model to obtain dynamic word vectors, and capture semantic features. These domain texts have many similar features to financial securities texts, providing guiding significance for entity extraction in the financial securities domain.

To our knowledge, research on NER in Chinese financial securities texts is still lacking. Therefore, considering the features, and challenges of Chinese financial securities texts, this paper also attempts to apply this multi-granularity feature extraction method to this field.

3. Method

This section describes the basic architecture of our proposed Chinese financial securities NER model, MFF-CNER. The model consists of four layers: The first layer is the BERT pre-training layer, which trains the model in an unsupervised manner on a large-scale unlabeled dataset to extract rich syntactic, and semantic features. This layer produces dynamic word embeddings with abundant language knowledge. The second layer is the feature extraction fusion layer, which is responsible for lower-level feature extraction. It utilizes BiLSTM to extract sentence-level contextual features, and then feeds the output of BiLSTM into IDCNN to further extract local features. This facilitates the fusion, and extraction of features. The third layer is the attention mechanism layer, where Multi-Head Attention is employed to calculate the weights of features, focusing on the ones that play a crucial role in classification. The fourth layer is the feature decoding layer, which utilizes a CRF model to decode the output of Attention, resulting in an optimal label sequence that captures the entities. The Chinese financial securities NER model is illustrated in Figure 1.

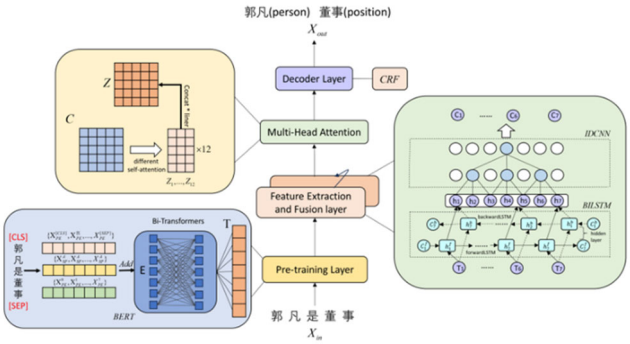

Figure 1. Framework of MFF-CNER.

3.1. Pre-training Layer

BERT is a deep, and bidirectional language representation model that is unsupervised, and utilized for pre-training [24]. Conventional approaches to word embedding, like Word2Vec [25], and GloVe [26], have a disadvantage in that identical words in distinct sentences have identical embeddings. However, BERT can tackle this problem. When compared with other pre-training language models like ELMo [27], and GPT [28], BERT has attained superior performance on NLP tasks [22].

The structure of BERT is shown in Figure 2, where E_1、E_2、…、E_n are the encoding representations of words, and they are obtained by training on a vast quantity of untagged data to derive word vector representations T_1、T_2、…、T_n. Here, Trm refers to the Transformer [29] structure.

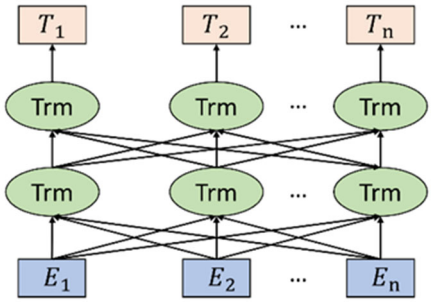

Figure 2. Structure of the BERT model.

The BERT pre-trained model primarily consists of a bidirectional Transformer architecture. Transformers utilize multi-head attention mechanisms, which can capture information from multiple dimensions to improve the representation ability of word embeddings. Consequently, the word vectors generated through BERT training are dynamic embeddings that can capture long-range sequential features.

As the attention mechanism fails to capture the positional information of words, the BERT architecture incorporates token embeddings, segment embeddings, and position embeddings in its input. The BERT model is trained through two tasks in parallel: Masked Language Model (MLM), and Next Sentence Prediction (NSP). The word embeddings can represent various character-level, word-level, sentence-level, and inter-sentence relations through this learning paradigm.

Pre-training financial securities text using BERT model can provide more contextual semantic knowledge to word embeddings, which can be fed into the feature extraction, and fusion layer to improve the feature extraction performance by BiLSTM, and IDCNN.

3.2. Feature Extraction and Fusion Layer

3.2.1. BiLSTM layer

The BiLSTM consists of two layers of LSTM in opposite directions, , and its advantage lies in its ability to capture both future , and past information. Figure 3 displays its configuration.

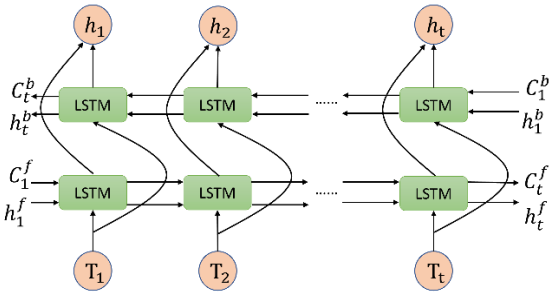

Figure 3. Structure of BiLSTM model.

The input sequence is denoted as $T=[T_1,T_2,...,T_n]$, which represents the output of a BERT pre-trained model. $H=[h_1,h_2,\ldots,h_n]$ represents the output sequence after passing through a BiLSTM model, where n denotes the sequence length. In the figure, x_t represents the input at time step t, h_t represents the output at time step t, , and h_t^f, h_t^b denote the forward , and backward output vectors of the LSTM, respectively. Furthermore, equation (1) is defined as the concatenation of the two vectors along the feature dimension, where $\oplus$ represents this operation.

$$h_t = h_t^f \oplus h_t^b . \tag{1}$$

The basic building block of BiLSTM is LSTM, a particular RNN model type. The model employs three gate controlling units, namely the input gate, forget gate, and output gate, to control the information that needs to be remembered, and forgotten at each time step. Because of the control provided by these three gates, LSTM can remember more extended sequence features, and solve the problems of vanishing, and exploding gradients that often occur during the training process of traditional RNN models. Figure 4 displays the structure of the model.

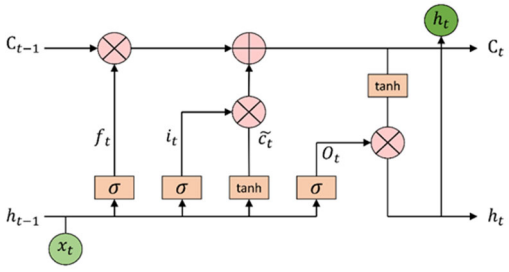

Figure 4. Structure of LSTM model.

The update of the three gate states is shown in Equations (2) to (7):

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f\right). \tag{2}$$

$$i_t = \sigma\left(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i\right). \tag{3}$$

$$o_t = \sigma\left(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o\right). \quad (4)$$

$$\tilde{c}_t = \tanh\left(W_{cx}x_t + W_{ch}h_{t-1} + b_c\right). \quad (5)$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t. \quad (6)$$

$$h_t = o_t \tanh\left(c_t\right). \quad (7)$$

f_t , i_t , and o_t respectively represent the states of the forget gate, input gate, , and output gate at time t . W is the weight matrix, which is learned through training. σ is the sigmoid activation function. b is the bias term, which is randomly initialized, and not trained. c_t is the cell state at time t , representing the LSTM neuron's memory information. h_t is the hidden state information at time t .

In financial securities text, many long texts have a high density of domain-specific terminology. By leveraging BiLSTM models, dependency relationships among words in lengthy texts can be captured, and features of the relationships between words can be extracted, thereby enhancing the accuracy of entity extraction.

3.2.2. IDCNN Layer

IDCNN contains several DCNN layers, whereas conventional CNN's final neurons can only acquire a limited fraction of knowledge from the input text. More convolutional layers must be added to acquire contextual information, leading to deeper networks with more parameters, and a higher risk of overfitting. DCNN can address this issue effectively by introducing a hyperparameter called Dilation Rate, which expands the convolutional kernel, and increases its receptive field. Figure 5 illustrates this concept.

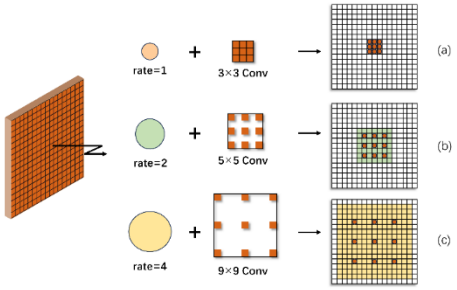

Figure 5. Dilated Convolution IDCNN Model.

Figure 5 illustrates the difference between regular, and dilated convolution, using a 3×3 kernel as an example. (a) The standard convolution is illustrated using a dilation rate, yielding a 3×3 receptive field. (b) Employing 2 dilation rate increases the receptive field to 5×5. (c) Doubling the dilation rate to 4 expands the receptive field to 15×15.

The internal structure of DCNN is shown in equations (8) to (10):

$$c_t^{(1)} = D_1^{(0)} h_t. \quad (8)$$

$$c_t^{(j)} = r\left(D_{2^j}^{(j-1)} c_t^{(j-1)}\right). \quad (9)$$

$$c_t^{(n+1)} = r(D_1^{(n)} c_t^{(n)}). \quad (10)$$

In this equation, $H = \left[h_1, h_2, \mathrm{B}\ , h_n\right]$ represents the output of BiLSTM, $D_\alpha^{(j)}$ is the dilation convolution at dilation distance α in layer j , $D_1^{(0)}$ is the dilation convolution at dilation distance 1 in the first layer. $c_t^{(j)}$ represents the feature obtained by the j -th layer of dilation convolution $(j>1)$; $r()$ denotes the ReLU activation function.

Through experiments, we selected the BiLSTM-IDCNN model for feature extraction, which outperformed other models, such as standalone BiLSTM or combinations of IDCNN with other models, and can better capture long-term dependencies in financial securities text by increasing the model's receptive field. For input text, IDCNN outputs the probabilities of each word corresponding to its respective labels.

3.3. Multi-Head Attention Layer

The multi-head attention is crucial in enhancing entity recognition accuracy by assigning weights to features extracted from BiLSTM-IDCNN feature vectors, accentuating essential features for classification, and attenuating less relevant features. Experimental results have demonstrated that incorporating attention mechanism can substantially enhance recognition performance.

The output of the multi-head attention layer is obtained by concatenating the outputs of multiple self-attention mechanisms, and applying a linear transformation. The formula is as follows:

$$\mathrm{Attention}\left(\mathrm{Q,\ K,\ V}\right) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (11)$$

$$MultiHead\left(Q,K,V\right) = \mathrm{Concat}\left(head_1, head_2, \mathrm{B}\ head_n\right)W. \quad (12)$$

$$head_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^v\right). \quad (13)$$

In the equation, Q, K, , and V represent the Query, Key, , and Value vectors. d_k is the dimension of the input vector. $head_i$ is the output of each self-attention mechanism.

3.4. Feature Decoder Layer

CRF is a commonly used model in the classic sequence labeling task, which considers the dependency between adjacent labels, and optimizes the label sequence from a global perspective, allowing it to consider all possible results when making predictions, and choose the outcome with the highest likelihood.

The fundamental operational mechanism of CRF involves computing the score function for the anticipated sequence given an input sequence $X = \left[x_1, x_2, \mathrm{B}\ , x_n\right]$, , and then determining the probability of $Y = \left[y_1, y_2, \mathrm{B}\ , y_n\right]$. Subsequently, using the likelihood maximization principle, the labeling sequence with the highest probability of the predicted sequence is selected as the output, given the output sequence corresponding to the input sequence $X = \left[x_1, x_2, \mathrm{B}\ , x_n\right]$.

Equation (14) reveals the score function for the forecasted sequence Y .

$$s(X,Y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i}. \quad (14)$$

Matrix A denotes the transition probability, where $A_{i,j}$ specifies the likelihood of the transition from label i to label j. Probability matrix P is output from the preceding layer, where $P_{i,j}$ indicates the probability that the i-th term is tagged with the j-th label.

The likelihood of the anticipated sequence Y is computed by means of Equation (15).

$$P(Y|X)=\frac{\exp(s(X,Y))}{\sum_{\tilde{Y}\in Y_X}s(X,\tilde{Y})}. \quad (15)$$

Y_x denotes the set of all conceivable label sequences, while $\tilde{Y}$ indicates the actual label sequence.

By applying a logarithmic function to both sides of equation (15), equation (16) presents the probability function for the anticipated sequence.

$$\ln(P(Y|X))=s(X,Y)-\ln\left(\sum_{\tilde{Y}\in Y_X}s(X,\tilde{Y})\right). \quad (16)$$

4. Experimental Results and Analysis

4.1. The Dataset

4.1.1. Dataset Gathering and Preparation

The text in this dataset is sourced from the announcements of Chinese A-share listed companies, industry research reports, and financial news between 2021, and 2022. The financial news text is obtained from Sina Finance (https://finance.sina.com.cn/), containing a wealth of information on intercompany relationships in A-share listed companies, such as related party transactions, and shareholding control. The announcements, and research reports of listed companies are sourced from the financial data platform Tushare (https://tushare.pro). These texts serve as crucial references for investor decision-making, and include substantial information about entities such as institutions, organizations, and individuals.

Through text clustering, and manual screening using domain expertise, the financial news articles were curated to retain only those relevant to changes in company ownership, executive management, and stock market developments related to A-share listed companies. Similarly, the announcements, and research reports were filtered to preserve texts relevant to specific industries, and individual stocks. As a result, a dataset consisting of 8,535 news texts, and 4,474 announcements, and research reports was obtained.

To improve the experimental effectiveness, preprocessing was performed on the text data, encompassing two steps. Firstly, special characters in the text that lack actual semantic meaning, such as "®", "■", and spaces, were removed. Secondly, any occurrences of traditional Chinese characters in the text were converted to simplified Chinese characters.

The Chinese financial securities dataset encompasses six categories of entities that influence security prices. The data was annotated in this experiment using the BIO [30] labeling method, with the Label-Studio(https://labelstud.io/) platform as the annotation tool. Illustrations of the annotations can be found in Figure 6, and Figure 7. The dataset was divided into training, validation, and test sets for this experiment, with an allocation ratio of 8:1:1.

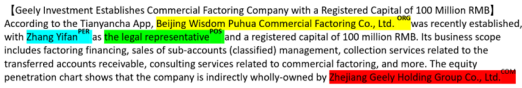

Figure 6. Labeling example of label-studio.

B-COM	I-COM	I-COM	I-COM	I-COM
北	京	智	慧	普
I-COM	I-COM	I-COM	I-COM	I-COM
华	商	业	保	理
I-COM	I-COM	I-COM	I-COM	O
有	限	公	司	成
O	O	B-POS	I-POS	I-POS
立	,	法	定	代
I-POS	I-POS	O	B-PER	I-PER
表	人	为	张	一
I-PER				
凡				

Figure 7. Example of BIO labeling.

The specific details of entity categories in the Chinese financial securities dataset are presented in Table 1.

Table 1. Statistics of Chinese Financial Securities Entities.

Label	NUM	Category	Entity Samples
PER	1564	Name of person	Zhang Yifan
COM	2239	Name of China A-share listed companies	CATL New Energy Technology Co., Ltd.
ORG	3286	Name of other social organizations	Zhejiang Provincial Government
POS	1244	The position of the person in the organization	Chairman
RISK	251	Risks to the company	Transaction risk
NOT	1201	Announcement issued by the company	Ant Technology Group Co., Ltd. undergoes industrial , and commercial changes

4.1.2. Other Datasets

In order to validate the applicability of the MFF-CNER model in the Chinese financial securities domain, we also applied the model along with five other baseline models to the Weibo [31, 32] dataset, and the MSRA [33] dataset, representing the social media, and news domains, respectively. Since the MSRA dataset does not have a validation set, the test set was divided into a 5:5 ratio. The basic information of these two datasets is presented in Table 2.

Table 2. Statistics of Weibo, and MSRA.

Dataset	Type	Train	Test	Dev
Weibo	Sentences	1.4k	0.27k	0.27k
	Char	73.8k	14.8k	14.5k
	Entities	1.89k	0.42k	0.39k
MSRA	Sentences	46.4k	4.4k	—
	Char	2169.9k	172.6k	—
	Entities	74.8k	6.2k	—

4.2. Experiment Settings

Our model was implemented in the PyTorch framework, , and the Adam optimizer was utilized to tune the parameters. The weight decay was set to 0.00005, and the GPU employed was the RTX A5000. The specific parameters for each model are presented in Table 3.

Table 3. Parameter settings of various methods.

Models	Parameters and Description	Models	Parameters and Description
BiLSTM-CRF	Max sequence length: 64 Bach size: 16 RNN layer: 1 RNN hidden dimension: 128 Learning rate: 0.0003 Dropout rate: 0.3 Epochs: 64	UIE	Model: uie-base Max sequence length: 512 Batch size: 16 Learning rate: 0.00001 Epochs: 20
IDCNN-CRF	Max sequence length: 64 Bach size: 16 Convolutional filters: 120 Learning rate: 0.0003 Dropout rate: 0.3 Epochs: 64	GlobalPoint	Encoder: chinese-bert-wwm Activation function: gelu Max sequence length: 128 Bach size: 16 Learning rate: 0.00002 Dropout rate: 0.1 Epochs: 5 Attention heads: 12 Hidden size: 768
BERT_CRF	Max sequence length: 64 Bach size: 16 Bert embedding dimension: 768 Learning rate: 0.0001 Dropout rate: 0.8 Epochs: 64	MFF-CNER	Max sequence length: 128 Bach size: 64 Bert embedding dimension: 768 RNN layer: 1 RNN hidden dimension: 128 Attention heads: 12 Convolutional filters: 120 Learning rate: 0.00003 Dropout rate: 0.3 Epochs: 48

4.3. Evaluation Indicators

The assessment of NER performance in this investigation utilized precision (P), recall (R), and F1-score measures, which are mathematically expressed as follows:

$$\mathrm{Precision}(P)=\frac{TP}{TP+FP}\times 100\%. \tag{17}$$

$$\mathrm{Recall}(R)=\frac{TP}{TP+FN}\times 100\%. \tag{18}$$

$$\mathrm{F1}=\frac{2\times P\times R}{P+R}\times 100\%. \tag{19}$$

TP denotes the scenarios in which the predicted , and actual values are both affirmative; FP corresponds to the scenarios in which the predicted value is positive while the actual value is negative; FN represents the scenarios in which the predicted value is negative while the actual value is positive.

4.4. Baseline Methods

BiLSTM-CRF [13]: This method is a commonly used model architecture for NER in the financial domain.

IDCNN-CRF [20]: The IDCNN network exhibits stronger capabilities in extracting local information than the BiLSTM network. The MFF-CNER method also utilizes this structure.

BERT-CRF: This method employs BERT as a pre-training model, and utilizes CRF for entity prediction decoding. It is currently a mainstream model frequently used as a benchmark for comparison.

GlobalPoint [34]: The model proposed by Su J et al. takes a global perspective to consider the starting position of entities, enabling unbiased recognition of both nested, and non-nested entities.

UIE [35]: The model Lu Y et al. proposed is a unified text-to-structure generation framework specifically designed for information extraction. It has achieved state-of-the-art entity extraction performance across multiple datasets.

4.5. Analysis of Experimental Results

The experimental results of these six models on different datasets are presented in Table 4.

Table 4. Experiment results of NER based on different models.

Models	My dataset			Weibo			MSRA		
	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%
BiLSTM-CRF	86.54	81.83	84.12	60.80	52.90	56.60	88.80	87.16	87.97
IDCNN-CRF	85.98	82.87	84.39	61.04	51.21	55.69	89.93	84.74	87.26
BERT-CRF	86.90	85.65	86.27	72.30	64.02	67.91	93.65	92.78	93.21
GlobalPointer	—	—	86.29	—	—	64.16	—	—	96.09
UIE	89.74	86.06	87.86	74.05	70.08	72.01	95.55	92.71	94.11
MFF-CNER	88.92	89.72	89.31	66.12	62.15	64.20	94.79	93.93	94.35

4.5.1. Comparison of different models

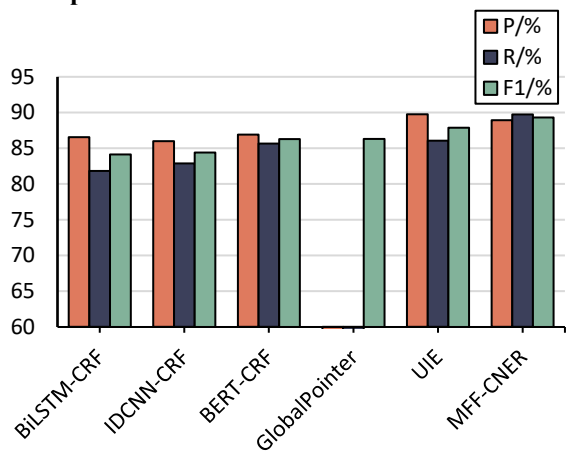

Figure 8. Experiment results of different models on My dataset.

From Figure 8, it can be observed that IDCNN, and BiLSTM exhibit similar performance in entity recognition. IDCNN shows a higher recall rate at 82.87%, while BiLSTM achieves a higher precision rate at 86.54%. The BERT-CRF model outperforms BiLSTM-CRF, and IDCNN-CRF in precision, recall, and F1 score, with a notable increase of 2.78% in recall compared to IDCNN. The reason behind this improvement lies in the fact that financial securities text contains dispersed entity information, long sentences, and complex syntax. Compared to BiLSTM, and IDCNN, BERT can learn more intricate semantic features.

GlobalPointer performs global normalization on financial securities text to address nested entity problems. Its overall performance is similar to that of the BERT-CRF model, with a slight improvement of 0.76% in precision but a decrease of 0.69% in recall. This indicates that financial semantic information is equally important in this domain as nested entity problems.

UIE, fine-tuned, and trained on training data, surpasses the performance of the previous four models, achieving precision, recall, and F1 scores of 89.74%, 86.06%, , and 87.86%, respectively. Although this model achieves state-of-the-art results on multiple datasets, a certain gap exists compared to MFF-CNER.

MFF-CNER achieves the best performance, where precision, recall, and F1 scores reach 88.92%, 89.72%, , and 89.31%, respectively. This demonstrates that the model can capture richer features. In entity extraction, the MFF-CNER model utilizes semantic features of text, and considers character-level context features, and local text features while incorporating attention mechanisms to enhance feature information. This model aligns with the textual, and entity features prevalent in this domain, ultimately achieving excellent performance in terms of overall effectiveness.

4.5.2. Performance of MFF-CNER

Table 5. Experimental results of MFF-CNER entities.

Entity	P/%	R/%	F1/%
PER	96.13	75.16	84.36
COM	94.58	91.21	92.97
ORG	77.51	78.57	78.03
POS	68.72	70.87	69.77
RISK	56.21	23.14	32.78
NOT	95.10	91.21	93.12

Table 5 presents the results of our proposed approach in recognizing various entity categories in the dataset. The table indicates that the F1 measure of the risk entity (32.78%) is the least, significantly inferior to other entities like company announcement (93.12%), company name (92.97%), and personal name (84.36%). The inferior performance of risk entity is due to its limited annotated samples, and the inherent intricacy in determining its boundaries, posing challenges for segmentation. For instance, the text "流动性紧张及债务风险 (liquidity stress , and debt risk)" includes two types of risks, liquidity risk , and debt risk, , and the former is difficult to identify due to ambiguous boundaries , and significant noise during the recognition process, resulting in difficulty in identification.

The F1 score of person position entity (69.77%) is also relatively low, with little difference between precision, and recall, due to the uneven distribution of this entity in the dataset. The entity "法定代表人 (legal representative)" appears more frequently than others, such as "总经理 (general manager)" , and "董事(director)", posing a challenge for the model to capture the distinctive features of these entities fully.

The precision of person name entity reaches 96.13%, while the recall is only 75.16%, with a precision much higher than the recall. This is because the features of this entity are relatively obvious, and easy to be captured by the model. However, the frequency of the same entity appearing in the text is relatively low, which often fails to recognize these entities when their features are not evident.

4.5.3. Analysis of domain applicability

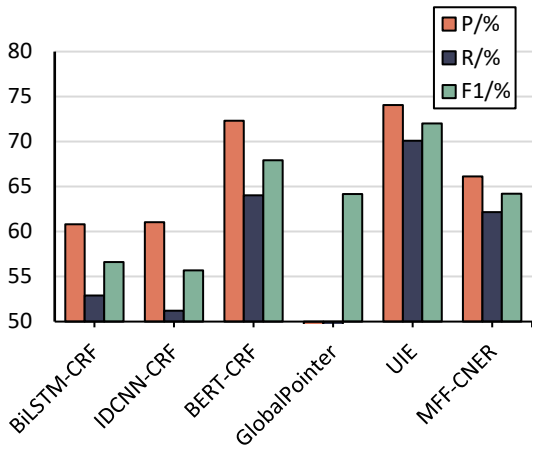

Figure 9. Experiment results of different models on Weibo.

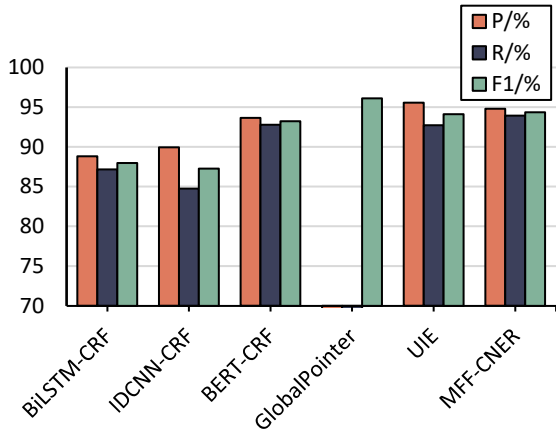

Figure 10. Experiment results of different models on MSRA.

To investigate the applicability of the MFF-CNER model in the Chinese financial securities domain, we compared these six models on two different domains: Weibo, which represents the social media domain, and MSRA, which represents the news domain.

On the Weibo dataset, the UIE model achieved the best performance with an F1 score of 72.01%. The GlobalPointer model ranked second with an F1 score of 64.16%, while the F1 score of the MFF-CNER model reached only 64.20%. Furthermore, both the precision, and recall of the MFF-CNER model were lower than those of UIE. On the MSRA dataset, the GlobalPointer model exhibited the best performance with an F1 score of 96.09%, while the F1 score of the MFF-CNER model was 94.35%, a mere 0.24% higher than UIE.

These experimental results indicate that the MFF-CNER model may not be suitable for other domains, and validate its applicability in the Chinese financial securities domain.

4.5.4. Ablation study

Table 6. Ablation of MFF-CNER. Baseline1 is BiLSTM-CRF.

	Baseline	BERT	IDCNN	Attention	P/%	R/%	F1/%
Model 1	√				86.54	81.83	84.12
Model 2	√	√			85.70	86.97	86.33
Model 3	√	√	√		87.80	86.04	86.91
Model 4	√	√	√	√	88.92	89.72	89.31

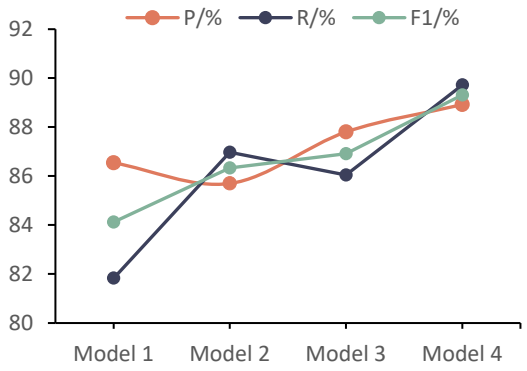

Figure 11. Changes in indicators of ablation experiments.

In order to investigate the contributions of different components in the MFF-CNER model, we conducted an ablation study on the constructed financial securities dataset. The results are presented in Table 6. Our findings are as follows:

1. Each component in MFF-CNER demonstrates positive effects. As shown in Figure 11, the F1 score consistently increases with the addition of components. However, on the Weibo dataset, BERT-CRF performs better than MFF-CNER (Table 5).

2. Attention plays a crucial role. Figure 11 shows a 2.4% increase in F1 score after incorporating Attention. This indicates that weighting multiple features through Attention is suitable for this domain.

3. BERT significantly enhances the model's ability to capture more entities. Figure 11 demonstrates a significant improvement in recall after incorporating BERT. This suggests that the model has learned more semantic features from the text, and become more sensitive to entities.

4. IDCNN effectively improves the precision of the model. A 2.1% increase in precision demonstrates the beneficial impact of capturing local features on entity recognition.

5. Conclusion

In this paper, we propose a NER method suited for the Chinese financial securities domain. Due to the longer boundaries, and greater diversity of expressions of entities in Chinese financial securities domain texts compared to general-domain texts [13], many distinguishing features are not easily captured. Therefore, we propose a multi-feature fusion method, MFF-CNER, to address this issue. The advantage of this model lies in its ability to extract features of different granularities from the text without encountering conflicts.

First, we conducted comparative experiments between the MFF-CNER model, and five baseline models. The results demonstrate that the MFF-CNER model achieves state-of-the-art performance. Next, we analyzed the recognition performance of the MFF-CNER model on different entity types in the domain. It was observed that there are still some challenges in identifying certain entities, such as risk entities, and position entities. The limited quantity of position entities in the dataset construction process contributes to this issue. The difficulty in delineating boundaries, and the presence of numerous noisy information account for the challenges associated with risk entities. Furthermore, we evaluated the performance of the MFF-CNER model in the domain-specific datasets of social media, and news, comparing it with other baseline models. The findings indicate that the MFF-CNER model achieves the best results only in the Chinese financial securities domain, thus confirming its applicability in this specific domain. Finally, we conducted ablation experiments on the MFF-CNER model based on the BiLSTM-CRF architecture. These experiments aimed to validate each component's positive contribution, and ensure no conflicts among the components. Notably, the BERT, and Attention modules were found to play a crucial role in significantly improving the model's performance.

To further enhance the accuracy of entity recognition in the Chinese financial securities domain , and address the potential applications in building financial securities knowledge graphs , and analyzing factors impacting stock prices, future research will focus on the following aspects: 1) Enriching the construction of domain-specific datasets by including more refined entity types; 2) Optimization of the MFF-CNER model to address the problem of not being able to delineate well the boundaries of certain entities (e.g., risks); 3) Much of the relevant information in this field is in the form of pictures , and tables, in order to further improve the performance of the model, more multimodal features are to be introduced to help entity recognition in the field of Chinese financial securities.

Acknowledgment

We thank Yanting Ji. This work was supported in part by a grant from the NNSF of China under Grant (No. 12271483).

References

[1] Sharnagat, R., Named entity recognition: A literature survey. *Center For Indian Language Technology* **2014**, 1-27.W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Jayakumar, H.; Krishnakumar, M. S.; Peddagopu, V. V. V.; Sridhar, R., RNN based question answer generation and ranking for financial documents using financial NER. *Sādhanā* **2020,** *45*, 1-10.

[3] Lamm, M.; Palomaki, J.; Alberti, C.; Andor, D.; Choi, E.; Soares, L. B.; Collins, M., Qed: A framework and dataset for explanations in question answering. Transactions of the Association for computational Linguistics 2021, 9, 790-806.

[4] Araújo, M.; Pereira, A.; Benevenuto, F., A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences* **2020,** *512*, 1078-1102.

[5] Rubino, R.; Fujita, A.; Marie, B. In *Error identification for machine translation with metric embedding and attention*, Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, 2021; pp 146-156.

[6] Bosselut, A.; Le Bras, R.; Choi, Y. In *Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering*, Proceedings of the AAAI conference on Artificial Intelligence, 2021; pp 4923-4931.

[7] Wang, W.; Xu, Y.; Du, C.; Chen, Y.; Wang, Y.; Wen, H., Data set and evaluation of automated construction of financial knowledge graph. *Data Intelligence* **2021,** *3* (3), 418-443.

[8] Li, J.; Sun, A.; Han, J.; Li, C., A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* **2020,** *34* (1), 50-70.

[9] Liu, P.; Guo, Y.; Wang, F.; Li, G., Chinese named entity recognition: The state of the art. *Neurocomputing* **2022,** *473*, 37-53.

[10] Asghari, M.; Sierra-Sosa, D.; Elmaghraby, A. S., BINER: A low-cost biomedical named entity recognition. *Information Sciences* **2022,** *602*, 184-200.

[11] Li, Y.; Wang, X.; Hui, L.; Zou, L.; Li, H.; Xu, L.; Liu, W., Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations. *JMIR Medical Informatics* **2020,** *8* (9), e19848.

[12] Trewartha, A.; Walker, N.; Huo, H.; Lee, S.; Cruse, K.; Dagdelen, J.; Dunn, A.; Persson, K. A.; Ceder, G.; Jain, A., Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **2022,** *3* (4), 100488.

[13] Qiurong, X.; Peng, Z.; Yifeng, L.; Qiwen, D., Research progress in Chinese named entity recognition in the financial field. *Journal of East China Normal University (Natural Science)* **2021,** *2021* (5), 1.

[14] Liu, Y.; Li, X.; Shi, J.; Zhang, L.; Li, J. In *Named entity recognition using a semi-supervised model based on bert and bootstrapping*, Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence: 5th China Conference, CCKS 2020, Nanchang, China, November 12–15, 2020, Revised Selected Papers, Springer: 2021; pp 54-63.

[15] Yuhan, L.; Changjian, L.; Ruifeng, X.; Wangda, L., Utilizing glyph feature and iterative learning for named entity recognition in finance text. *Journal of Chinese Information Processing* **2020,** *34* (11), 74-83.

[16] Huang, Z.; Xu, W.; Yu, K., Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* **2015**.

[17] Shen, J.; Li, F.; Xu, F.; Uszkoreit, H., Recognition of chinese organization names and abbreviations. *Journal of Chinese Information Processing* **2007,** *21* (6), 17-21.

[18] Wang, S.; Xu, R.; Liu, B.; Gui, L.; Zhou, Y. In *Financial named entity recognition based on conditional random fields and information entropy*, 2014 international conference on machine learning and cybernetics, IEEE: 2014; pp 838-843.

[19] Yang, T.; Jiang, D.; Shi, S.; Zhan, S.; Zhuo, L.; Yin, Y.; Liang, Z. In *Chinese data extraction and named entity recognition*, 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), IEEE: 2020; pp 105-109.

[20] Strubell, E.; Verga, P.; Belanger, D.; McCallum, A., Fast and accurate entity recognition with iterated dilated convolutions. *arXiv preprint arXiv:1702.02098* **2017**.

[21] Chang, Y.; Kong, L.; Jia, K.; Meng, Q. In *Chinese named entity recognition method based on BERT*, 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), IEEE: 2021; pp 294-299.

[22] Li, Z.; Yun, H.; Guo, Z.; Qi, J. In *Medical Named Entity Recognition Based on Multi Feature Fusion of BERT*, Proceedings of the 4th International Conference on Big Data Technologies, 2021; pp 86-91.

[23] Zhang, Y.; Wang, S.; He, B.; Ye, P.; Li, K., Named entity recognition method of elementary mathematical text based on BERT. *Journal of Computer Applications* **2022,** *42* (2), 433.

[24] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018.

[25] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J., Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* **2013**.

[26] Pennington, J.; Socher, R.; Manning, C. D. In Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014; pp 1532-1543.

[27] Sarzynska-Wawer, J.; Wawer, A.; Pawlak, A.; Szymanowska, J.; Stefaniak, I.; Jarkiewicz, M.; Okruszek, L., Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research 2021, 304, 114135.

[28] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I., Improving language understanding by generative pre-training. **2018**.

[29] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I., Attention is all you need. *Advances in neural information processing systems* **2017,** *30*.

[30] Sang, E. F.; Veenstra, J., Representing text chunks. *arXiv preprint cs/9907006* **1999**.

[31] Peng, N.; Dredze, M. In *Named entity recognition for chinese social media with jointly trained embeddings*, Proceedings of the 2015 conference on empirical methods in natural language processing, 2015; pp 548-554.

[32] Peng, N.; Dredze, M., Improving named entity recognition for chinese social media with word segmentation representation learning. arXiv preprint arXiv:1603.00786 2016.

[33] Levow, G.-A. In The third international Chinese language processing bakeoff: Word segmentation and named entity recognition, Proceedings of the Fifth SIGHAN workshop on Chinese language processing, 2006; pp 108-117.

[34] Su, J.; Murtadha, A.; Pan, S.; Hou, J.; Sun, J.; Huang, W.; Wen, B.; Liu, Y., Global Pointer: Novel Efficient Span-based Approach for Named Entity Recognition. arXiv preprint arXiv:2208.03054 2022.

[35] Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; Wu, H., Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277* **2022**.